Inferring Super-Resolution Depth from a Moving Light-Source Enhanced RGB-D Sensor: A Variational Approach

Lu Sang, Bjoern Haefner, Daniel Cremers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1-10

A novel approach towards depth map super-resolution using multi-view uncalibrated photometric stereo is presented. Practically, an LED light source is attached to a commodity RGB-D sensor and is used to capture objects from multiple viewpoints with unknown motion. This non-static camera-to-object setup is described with a nonconvex variational approach such that no calibration on lighting or camera motion is require due to the formulation of an end-to-end joint optimization problem. Solving the proposed variational model results in high resolution depth, reflectance and camera estimates, as we show on challenging synthetic and real-world datasets.

****

Unsupervised Learning of Camera Pose with Compositional Re-estimation

Seyed shahabeddin Nabavi, Mehrdad Hosseinzadeh, Ramin Fahimi, Yang Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 11-20

We consider the problem of unsupervised camera pose estimation. Given an input video sequence, our goal is to estimate the camera pose (i.e. the camera motion) between consecutive frames. Traditionally, this problem is tackled by placing strict constraints on the transformation vector or by incorporating optical flow through a complex pipeline. We propose an alternative approach that utilizes a compositional re-estimation process for camera pose estimation. Given an input, we first estimate a depth map. Our method then iteratively estimates the camera motion based on the estimated depth map. Our approach significantly improves the predicted camera motion both quantitatively and visually. Furthermore, the re-estimation resolves the problem of out-of-boundaries pixels in a novel and simple way. Another advantage of our approach is that it is adaptable to other camera pose estimation approaches. Experimental analysis on KITTI benchmark dataset demonstrates that our method outperforms existing state-of-the-art approaches in unsupervised camera ego-motion estimation.

****

Blended Convolution and Synthesis for Efficient Discrimination of 3D Shapes

Sameera Ramasinghe, Salman Khan, Nick Barnes, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 21-31

Existing models for shape analysis directly learn feature representations on 3D point clouds. We argue that 3D point clouds are highly redundant and hold irregular (permutation-invariant) structure, which makes it difficult to achieve inter-class discrimination efficiently. In this paper, we propose a two-pronged solution to this problem that is seamlessly integrated in a single blended convolution and synthesis layer. This fully differentiable layer performs two critical tasks in succession. In the first step, it projects the input 3D point clouds into a latent 3D space to synthesize a highly compact and inter-class discriminative point cloud representation. Since, 3D point clouds do not follow a Euclidean topology, standard 2/3D convolutional neural networks offer limited representation capability. Therefore, in the second step, we propose a novel 3D convolution operator functioning inside the unit ball to extract useful volumetric features. We derive formulae to achieve both translation and rotation of our novel convolution kernels. Finally, using the proposed techniques we present an extremely light-weight, end-to-end architecture that achieves compelling results on 3D shape recognition and retrieval.

****

A Multi-Scale Guided Cascade Hourglass Network for Depth Completion

Ang Li, Zejian Yuan, Yonggen Ling, Wanchao Chi, shenghao zhang, Chong Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 32-40

Depth completion, a task to estimate the dense depth map from sparse measurement under the guidance from the high-resolution image, is essential to many compute r vision applications. Most previous methods building on fully convolutional networks can not handle diverse patterns in the depth map efficiently and effectively. We propose a multi-scale guided cascade hourglass network to tackle this problem. Structures at different levels are captured by specialized hourglasses in the cascade network with sparse inputs in various sizes. An encoder extracts multiscale features from color image to provide deep guidance for all the hourglasses. A multi-scale training strategy further activates the effect of cascade stages. With the role of each sub-module divided explicitly, we can implement components with simple architectures. Extensive experiments show that our lightweight model achieves competitive results compared with state-of-the-art in KITTI depth completion benchmark, with low complexity in run-time.

****

Silhouette Guided Point Cloud Reconstruction beyond Occlusion
Chuhang Zou, Derek Hoiem; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 41-50

One major challenge in 3D reconstruction is to infer the complete shape geometry from partial foreground occlusions. In this paper, we propose a method to reconstruct the complete 3D shape of an object from a single RGB image, with robustness to occlusion. Given the image and a silhouette of the visible region, our approach completes the silhouette of the occluded region and then generates a point cloud. We show improvements for reconstruction of non-occluded and partially occluded objects by providing the predicted complete silhouette as guidance. We also improve state-of-the-art for 3D shape prediction with a 2D reprojection loss from multiple synthetic views and a surface-based smoothing and refinement step. Experiments demonstrate the efficacy of our approach both quantitatively and qualitatively on synthetic and real scene datasets.

****

Non-Rigid Structure from Motion: Prior-Free Factorization Method Revisited
Suryansh Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 51-60

A simple prior free factorization algorithm [??] is quite often cited work in the field of Non-Rigid Structure from Motion (NRSfM). The benefit of this work lies in its simplicity of implementation, strong theoretical justification to the motion and structure estimation, and its invincible originality. Despite this, the prevailing view is, that it performs exceedingly inferior to other methods on several benchmark datasets [??]. However, our subtle investigation provides some empirical statistics which made us think against such views. The statistical results we obtained supersedes Dai \it et al. [??] originally reported results on the benchmark datasets by a significant margin under some elementary changes in their core algorithmic idea [??]. Now, these results not only exposes some unrevealed areas for research in NRSfM but also give rise to new mathematical challenges for NRSfM researchers. We argue that by properly utilizing the well-established assumptions about a non-rigidly deforming shape i.e, it deforms smoothly over frames [??] and it spans a low-rank space, the simple prior-free idea can provide results which is comparable to the best available algorithms. In this paper, we explore some of the hidden intricacies missed by Dai \it et. al.   work [??] and how some elementary measures and modifications can enhance its performance, as high as approx. 18% on the benchmark dataset. The improved performance is justified and empirically verified by extensive experiments on several datasets. We believe our work has both practical and theoretical importance for the development of better NRSfM algorithms.

****

PointGrow: Autoregressively Learned Point Cloud Generation with Self-Attention
Yongbin Sun, Yue Wang, Ziwei Liu, Joshua Siegel, Sanjay Sarma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 61-70

Generating 3D point clouds is challenging yet highly desired. This work presents a novel autoregressive model, PointGrow, which can generate diverse and realistic point cloud samples from scratch or conditioned on semantic contexts. This model operates recurrently, with each point sampled according to a conditional dis tribution given its previously-generated points, allowing inter-point correlations to be well-exploited and 3D shape generative processes to be better interpreted. Since point cloud object shapes are typically encoded by long-range dependencies, we augment our model with dedicated self-attention modules to capture such relations. Extensive evaluations show that PointGrow achieves satisfying performance on both unconditional and conditional point cloud generation tasks, with respect to realism and diversity. Several important applications, such as unsupervised feature learning and shape arithmetic operations, are also demonstrated.

****

Depth Completion via Deep Basis Fitting
Chao Qu, Ty Nguyen, Camillo Taylor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 71-80

In this paper we consider the task of image-guided depth completion where our system must infer the depth at every pixel of an input image based on the image content and a sparse set of depth measurements. We propose a novel approach that builds upon the strengths of modern deep learning techniques and classical fitting algorithms and significantly improves performance. The proposed method replaces the final 1-by-1 convolutional layer employed in most depth completion networks with a least squares fitting module which computes weights by fitting the implicit depth bases to the given sparse depth measurements. In addition, we show how our proposed method can be naturally extended to a multi-scale formulation for improved self-supervised training. We demonstrate through extensive experiments on various datasets that our approach achieves consistent improvements over a state-of-the-art baseline method with minimal computational overhead.

****

High Accuracy Face Geometry Capture using a Smartphone Video
Shubham Agrawal, Anuj Pahuja, Simon Lucey; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 81-90

What's the most accurate 3D model of your face you can obtain while sitting at your desk? We attempt to answer this question in our work. High fidelity face reconstructions have so far been limited to either studio settings or through expensive 3D scanners. On the other hand, unconstrained reconstruction methods are typically limited by low-capacity models. Our method reconstructs accurate face geometry of a subject using a video shot from a smartphone in an unconstrained environment. Our approach takes advantage of recent advances in visual SLAM, keypoint detection, and object detection to improve accuracy and robustness. By not being constrained to a model subspace, our reconstructed meshes capture important details while being robust to noise and being topologically consistent. Our evaluations show that our method outperforms current single and multi-view baselines by a significant margin, both in terms of geometric accuracy and in capturing person-specific details important for making realistic looking models.

****

FlowNet3D++: Geometric Losses For Deep Scene Flow Estimation
Zirui Wang, Shuda Li, Henry Howard-Jenkins, Victor Prisacariu, Min Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 91-98

We present FlowNet3D++, a deep scene flow estimation network. Inspired by classical methods, FlowNet3D++ incorporates geometric constraints in the form of point-toplane distance and angular alignment between individual vectors in the flow field, into FlowNet3D. We demonstrate that the addition of these geometric loss terms improves the previous state-of-art FlowNet3D accuracy from 57.85% to 63.43%. To further demonstrate the effectiveness of our geometric constraints, we propose a benchmark for flow estimation on the task of dynamic 3D reconstruction, thus providing a more holistic and practical measure of performance than the breakdown of individual metrics previously used to evaluate scene flow. This is made possible through the contribution of a novel pipeline to integrate point-based scene flow predictions into a global dense volume. FlowNet3D++ achieves up to a 15.0% reduction in reconstruction error over FlowNet3D, and up to a 35.2% improvement over KillingFusion alone. We will release our scene flow estimation code later.

****

Style Transfer for Light Field Photography
David Hart, Bryan Morse, Jessica Greenland; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 99-108

As light field images continue to increase in use and application, it becomes necessary to adapt existing image processing methods to this unique form of photography. In this paper we explore methods for applying neural style transfer to light field images. Feed-forward style transfer networks provide fast, high-quality results for monocular images, but no such networks exist for full light field images. Because of the size of these images, current light field data sets are small and are insufficient for training purely feed-forward style-transfer networks from scratch. Thus, it is necessary to adapt existing monocular style transfer networks in a way that allows for the stylization of each view of the light field while maintaining visual consistencies between views. To do this, we first generate disparity maps for each view given a single depth image for the light field. Then in a fashion similar to neural stylization of stereo images, we use disparity maps to enforce a consistency loss between views and to warp feature maps during the feed forward stylization. Unlike previous work, however, light fields have too many views to train a purely feed-forward network that can stylize the entire light field with angular consistency. Instead, the proposed method uses an iterative optimization for each view of a single light field image that backpropagates the consistency loss through the network. Thus, the network architecture allows for the incorporation of pre-trained fast monocular stylization network while avoiding the need for a large light field training set.

****

Fourier Based Pre-Processing For Seeing Through Water
Jerin Geo James, Ajit Rajwade; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 109-117

Consider a scene submerged underneath a fluctuating water surface. Images of such a scene, when acquired from a camera in the air, exhibit significant spatial distortions. In this paper, we present a novel, computationally efficient pre-processing algorithm to correct a significant amount (~ 50%) of apparent distortion present in video sequences of such a scene. We demonstrate that when the partially restored video output from this stage is given as input to other methods, it significantly improves their performance. This algorithm involves (i) tracking a small number N of salient feature points across the T frames to yield point-trajectories \ \boldsymbol q_i \triangleq \ (x_ it ,y_ it)\ _ t=1 ^T\ _ i=1 ^N, and (ii) using the point-trajectories to infer the deformations at other non-tracked points in every frame. A Fourier decomposition of the N trajectories, followed by a novel Fourier phase-interpolation step, is used to infer deformations at all other points. Our method exploits the inherent spatio-temporal characteristics of the fluctuating water surface to correct non-rigid deformations to a very large extent. The source code, datasets and supplemental material can be accessed at [1], [2].

****

DeOccNet: Learning to See Through Foreground Occlusions in Light Fields
Yingqian Wang, Tianhao Wu, Jungang Yang, Longguang Wang, Wei An, Yulan Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 118-127

Background objects occluded in some views of a light field (LF) camera can be seen by other views. Consequently, occluded surfaces are possible to be reconstructed from LF images. In this paper, we handle the LF de-occlusion (LF-DeOcc) problem using a deep encoder-decoder network (namely, DeOccNet). In our method, sub-aperture images (SAIs) are first given to the encoder to incorporate both spatial and angular information. The encoded representations are then used by the decoder to render an occlusion-free center-view SAI. To the best of our knowledge, DeOccNet is the first deep learning-based LF-DeOcc method. To handle the insufficiency of training data, we propose an LF synthesis approach to embed selected occlusion masks into existing LF images. Besides, several synthetic and real-world LFs are developed for performance evaluation. Experimental results show that, after training on the generated data, our DeOccNet can effectively remove foreground occlusions and achieves superior performance as compared to other state-of-the-art methods. Source codes are available at: https://github.com/YingqianWang/DeOccNet.

****

Appearance and Shape from Water Reflection

Ryo Kawahara, Meng-Yu Kuo, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 128-136

This paper introduces single-image geometric and appearance reconstruction from water reflection photography, i.e., images capturing direct and water-reflected real-world scenes. Water reflection offers an additional viewpoint to the direct sight, collectively forming a stereo pair. The water-reflected scene, however, includes internally scattered and reflected environmental illumination in addition to the scene radiance, which precludes direct stereo matching. We derive a principled iterative method that disentangles this scene radiometry and geometry for reconstructing 3D scene structure as well as its high-dynamic range appearance. In the presence of waves, we simultaneously recover the wave geometry as surface normal perturbations of the water surface. Most important, we show that the water reflection enables calibration of the camera. In other words, for the first time, we show that capturing a direct and water-reflected scene in a single exposure forms a self-calibrating HDR catadioptric stereo camera. We demonstrate our method on a number of images taken in the wild. The results demonstrate a new means for leveraging this accidental catadioptric camera.

****

An Extended Exposure Fusion and its Application to Single Image Contrast Enhancement

Charles Hessel, Jean-Michel Morel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 137-146

Exposure Fusion is a high dynamic range imaging technique fusing a bracketed exposure sequence into a high quality image. In this paper, we provide a refined version resolving its out-of-range artifact and its low-frequency halo. It improves on the original Exposure Fusion by augmenting contrast in all image parts. Furthermore, we extend this algorithm to single exposure images, thereby turning it into a competitive contrast enhancement operator. To do so, bracketed images are first simulated from a single input image and then fused by the new version of Exposure Fusion. The resulting algorithm competes with state of the art image enhancement methods.

****

Online Lens Motion Smoothing for Video Autofocus

Abdullah Abuolaim, Michael Brown; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 147-155

Autofocus (AF) is the process of moving the camera's lens such that desired scene content is in focus.   AF for single image capture is a well-studied research topic and most modern cameras have hardware support that allows quick lens movements to optimize image sharpness. How to best perform AF for video is less clear. Conventional wisdom would suggest that each temporal frame should be as sharp as possible. However, unlike single image capture, the effects of the lens movement is visible in the captured video. As a result, there are two parameters to consider in AF for video: sharpness and lens movement. In this paper, we show that users preferred videos with smooth lens movement, even if it results in less overall sharpness.   Based on this observation, we propose two novel AF algorithms for video that strive for both smooth lens movement and sharp scene content. Specifically, we introduce (1) a bidirectional long short-term memory (BLSTM) module trained on smooth lens trajectories and (2) a simple weighted moving average (WMA) method that factors in prior lens motion. Both of these methods have demonstrated excellent results in terms of reducing lens movements (up to 64% reduction) without greatly affecting the sharpness (less than 5.2% change in sharpness). Moreover, videos produced using our methods are more preferred by users over conventional AF that aims only for maximizing sharpness.

****
Fast Image Reconstruction with an Event Camera
Cedric Scheerlinck, Henri Rebecq, Daniel Gehrig, Nick Barnes, Robert Mahony, Davide Scaramuzza; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 156-163
Event cameras are powerful new sensors able to capture high dynamic range with microsecond temporal resolution and no motion blur. Their strength is detecting brightness changes (called events) rather than capturing direct brightness images; however, algorithms can be used to convert events into usable image representations for applications such as classification. Previous works rely on hand-crafted spatial and temporal smoothing techniques to reconstruct images from events. State-of-the-art video reconstruction has recently been achieved using neural networks that are large (10M parameters) and computationally expensive, requiring 30ms for a forward-pass at 640 x 480 resolution on a modern GPU. We propose a novel neural network architecture for video reconstruction from events that is smaller (38k vs. 10M parameters) and faster (10ms vs. 30ms) than state-of-the-art with minimal impact to performance. Videos and Datasets: https://cedric-scheerlinck.github.io/firenet
****
Self-Guided Novel View Synthesis via Elastic Displacement Network
Yicun Liu, Jiawei Zhang, Ye Ma, Jimmy Ren; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 164-173
Synthesizing a novel view from different viewpoints has been an essential problem in 3D vision. Among a variety of view synthesis tasks, single image based view synthesis is particularly challenging. Recent works address this problem by a fixed number of image planes of discrete disparities, which tend to generate structurally inconsistent results on wide-baseline, scene-complicated datasets such as KITTI. In this paper, we propose the Self-Guided Elastic Displacement Network (SG-EDN), which explicitly models the geometric transformation by a novel non-discrete scene representation called layered displacement maps (LDM). To generate realistic views, we exploit the positional characteristics of the displacement maps and design a multi-scale structural pyramid for self-guided filtering on the displacement maps. To optimize efficiency and scene-adaptivity, we allow the effective range of each displacement map to be elastic, with fully learnable parameters. Experimental results confirm that our framework outperforms existing methods in both quantitative and qualitative tests.
****
On Scene Flow Computation of Gas Structures with Optical Gas Imaging Cameras
Johannes Rangel, Robert Schmoll, Andreas Kroll; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 174-182
Gas leak inspection and gas leak quantification are nowadays of high relevance within the oil and gas industry as well as in many other industrial sectors. This has been driven by safety-related issues, economic losses and the considerable climate impact caused by such unwanted gas releases. Due to the latter, the efforts for developing new and more reliable measurement techniques for detecting and quantifying greenhouse gases such as methane have increased in the recent years. In this work, a stereo camera system based on optical gas imaging cameras is used for computing dense 3D velocity information, i.e. scene flow, of escaping gas structures. Here, the optical flow, the disparity and the disparity change in likely gas image regions are computed utilizing classical variational methods. The accuracy of the applied methods and their applicability under real conditions in a biogas plant are characterized and tested. The results show that the recovered 3D gas velocity field per camera frame approaches the average 3D velocity field of the measured gas structure. The accuracy of the used method is affected, among others, when the imaged gas structures exhibit a low contrast.
****
Fast Deep Stereo with 2D Convolutional Processing of Cost Signatures
Kyle Yee, Ayan Chakrabarti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 183-191
Modern neural network-based algorithms are able to produce highly accurate depth

estimates from stereo image pairs, nearly matching the reliability of measurements from more expensive depth sensors. However, this accuracy comes with a higher computational cost since these methods use network architectures designed to compute and process matching scores across all candidate matches at all locations, with floating point computations repeated across a match volume with dimensions corresponding to both space and disparity. This leads to longer running times to process each image pair, making them impractical for real-time use in robots and autonomous vehicles. We propose a new stereo algorithm that employs a significantly more efficient network architecture. Our method builds an initial match cost volume using traditional matching costs that are fast to compute, and trains a network to estimate disparity from this volume. Crucially, our network only employs per-pixel and two-dimensional convolution operations: to summarize the local match information at each location as a low-dimensional feature vector, and to spatially process these "cost-signature" features to produce a dense disparity map. Experimental results on KITTI show that our method delivers competitive accuracy at significantly higher speeds---running at 48 frames per second on a modern GPU.
****

Triple-SGM: Stereo Processing using Semi-Global Matching with Cost Fusion
Jan Kallwies, Torsten Engler, Bianca Forkel, Hans-Joachim Wuensche; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 192-200

In this work, we propose an extension of the Semi-Global Matching framework for three images from a stereo rig consisting of a horizontal and vertical camera pair. After calculating the matching costs separately for both image pairs, these are merged at cost level using cubic spline interpolation. For cost values near the left/bottom image boundaries, we propose an advanced weighting strategy. Subsequently, the fused matching can be used directly for the cost aggregation and disparity estimation. The benefits of the proposed fusion strategy are demonstrated by an evaluation based on synthetic and real-world data. To encourage further comparisons on triple stereo algorithms, the dataset used for evaluation is made publicly available.
****

Optimizing Through Learned Errors for Accurate Sports Field Registration
Wei Jiang, Juan Camilo Gamboa Higuera, Baptiste Angles, Weiwei Sun, Mehrsan Javan, Kwang Moo Yi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 201-210

We propose an optimization-based framework to register sports field templates onto broadcast videos. For accurate registration we go beyond the prevalent feed-forward paradigm. Instead, we propose to train a deep network that regresses the registration error, and then register images by finding the registration parameters that minimize the regressed error. We demonstrate the effectiveness of our method by applying it to real-world sports broadcast videos, outperforming the state of the art. We further apply our method on a synthetic toy example and demonstrate that our method brings significant gains even when the problem is simplified and unlimited training data is available.
****

Reference Grid-assisted Network for 3D Point Signature Learning from Point Clouds
Jing Zhu, Yi Fang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 211-220

Learning a robust 3D point signature from point clouds is an interesting but challenging task in the computer vision field due to the irregular and unordered structure characteristics of the point cloud data. In this paper, we propose to learn a 3D point signature by exploring the implicit relation between keypoints and their neighbors (grouped as patches) among the given scene point clouds. Specially, we design a uniform reference grid to represent the raw relation between each keypoint and its neighbors from the raw point clouds. In order to learn a 3D point signature gradually from a smaller perceptive region to a larger area, we create a novel framework with a MLP-based unit feature network and a 3D CNN-bas ed grid feature network. Specifically, the unit feature network aims to dig the connections from points fallen into the same unit of the reference grid, while the grid feature network is used to discover the grid-wise relations across the whole reference grid with concatenation of the learned unit-wise features. Moreover, we introduce an MLP-based attention network upon the unit feature network to enhance the discriminative ability of our learned 3D point signature. All the components in our proposed model are implemented as siamese ones to better tackle the classic keypoint matching and geometric registration problems. Our proposed 3D point signature learning approach achieves superior performance over other state-of-the-art methods on keypoint matching and geometric registration on the real-world scenes datasets, e.g. SUN3D, 7-scenes and the synthetic scan augmented scenes in ICL-NUIM dataset. More importantly, our learned 3D point signature successfully handles the point cloud fragment alignment challenges by producing correct transformations with RANSAC algorithm.

****

Stable Intrinsic Auto-Calibration from Fundamental Matrices of Devices with Uncorrelated Camera Parameters

Torben Fetzer, Gerd Reis, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 221-230

Auto-Calibration is an important task in computer vision and is necessary for many visual applications. Methods like photogrammetry, depth map estimation, metrology, augmented/mixed reality or odometry are strongly dependent on well calibrated devices. While classical calibration relies on tools like checkerboards or additional scene information, auto-calibration only takes epipolar relations into account. Classical calibration is often impractical, tends to de-adjust over time and distributes the error over the entire, limited working volume. Auto-calibration, on the other hand, does not require any information other than the image content itself, has a virtually unlimited working range and usually achieves highest accuracy at the objects' surfaces. Unfortunately, auto-calibration methods are sensitive to errors in the fundamental matrix and need good initialization to converge to the global solution. In practice this leads to difficulties if optical parameters like principal point or focal length are unconstrained. In such situations, even state-of-the-art auto-calibration methods tend to diverge and do not yield a valid calibration. This work assesses reasons for this behavior, in particular for the initialization method of Bougnoux [3] and Lourakis' state-of-the-art auto-calibration method [21]. Based on the analysis, a more stable method is proposed. A continuous and smooth energy functional is introduced, providing superior convergence properties. I.e. it can not diverge, converges faster, and has a significantly enlarged convergence region with respect to the global minimum. Finally, a thorough evaluation has been conducted and a detailed comparison with the state of the art is presented.

****

Deep Image Blending

Lingzhi Zhang, Tarmily Wen, Jianbo Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 231-240

Image composition is an important operation to create visual content. Among image composition tasks, image blending aims to seamlessly blend an object from a source image onto a target image with lightly mask adjustment. A popular approach is Poisson image blending, which enforces the gradient domain smoothness in the composite image. However, this approach only considers the boundary pixels of target image, and thus can not adapt to texture of target image. In addition, the colors of the target image often seep through the original source object too much causing a significant loss of content of the source object. We propose a Poisson blending loss that achieves the same purpose of Poisson image blending. In addition, we jointly optimize the proposed Poisson blending loss as well as the style and content loss computed from a deep network, and reconstruct the blending region by iteratively updating the pixels using the L-BFGS solver. In the blending image, we not only smooth out gradient domain of the blending boundary but also add consistent texture into the blending region. User studies show that our method can outperform strong baselines as well as state-of-the-art approaches whe n placing objects onto both paintings and real-world images.

****

Cross-Domain Face Synthesis using a Controllable GAN

Fania Mokhayeri, Kaveh Kamali, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 252-260

The performance of face recognition (FR) systems for video surveillance has been shown to improve when the design data is augmented through synthetic face generation. This is true, for instance, with pair-wise matchers (e.g., deep Siamese networks) that rely on a reference gallery, typically with one still image per individual. However, generating synthetic images based on stills may not improve performance during operations due to the domain shift w.r.t. the target domain. Moreover, despite the emergence of Generative Adversarial Networks (GANs) for realistic synthetic generation, it is often difficult to control the conditions under which synthetic faces are generated. In this paper, a cross-domain face synthesis approach is proposed that integrates a new Controllable GAN (C-GAN). It employs an off-the-shelf 3D face model as a simulator to generate facial images under various poses. The simulated images and noise are input to the C-GAN for realism refinement. It relies on an additional adversarial game as a third player to preserve the identity and specific facial attributes of the refined images. This allows generating realistic synthetic face images that reflect capture conditions in the target domain, while controlling the GAN output such that faces may be generated under desired pose conditions. Experiments were performed using videos from the Chokepoint and COX-S2V datasets, and a deep Siamese network for FR with a single reference still per person. Results indicate that the proposed approach can provide a higher level of accuracy compared to state-of-the-art approaches for synthetic data augmentation.

****

Does Face Recognition Accuracy Get Better With Age? Deep Face Matchers Say No

Vitor Albiero, Kevin Bowyer, Kushal Vangara, Michael King; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 261-269

Previous studies generally agree that face recognition accuracy is higher for older persons than for younger persons. But most previous studies were before the wave of deep learning matchers, and most considered accuracy only in terms of the verification rate for genuine pairs. This paper investigates accuracy for age groups 16-29, 30-49 and 50-70, using three modern deep CNN matchers, and considers differences in the impostor and genuine distributions as well as verification rates and ROC curves. We find thataccuracy is lower for older persons and higher for younger persons. In contrast, a pre deep learning matcher on the same dataset shows the traditional result of higher accuracy for older persons, although its overall accuracy is much lower than that of the deep learning matchers. Comparing the impostor and genuine distributions, we conclude that impostor scores have a larger effect than genuine scores in causing lower accuracy for the older age group. We also investigate the effects of training data across the age groups. Our results show that fine-tuning the deep CNN models on additional images of older persons actually lowers accuracy for the older age group. Also, we fine-tune and train from scratch two models using age-balanced training datasets, and these results also show lower accuracy for older age group. These results argue that the lower accuracy for the older age group is not due to imbalance in the original training data.

****

Offset Calibration for Appearance-Based Gaze Estimation via Gaze Decomposition

Zhaokang Chen, Bertram Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 270-279

Appearance-based gaze estimation provides relatively unconstrained gaze tracking. However, subject-independent models achieve limited accuracy partly due to individual variations. To improve estimation, we propose a gaze decomposition method that enables low complexity calibration, i.e., using calibration data collected when subjects view only one or a few gaze targets and the number of images per gaze target is small. Lowering the complexity of calibration makes it more conv enient and less time-consuming for the user, and more widely applicable. Motivated by our finding that the inter-subject squared bias exceeds the intra-subject variance for a subject-independent estimator, we decompose the gaze estimate into the sum of a subject-independent term estimated from the input image by a deep convolutional network and a subject-dependent bias term. During training, both the weights of the deep network and the bias terms are estimated. During testing, if no calibration data is available, we can set the bias term to zero. Otherwise, the bias term can be estimated from images of the subject gazing at known gaze targets. Experimental results on three datasets show that without calibration, our method outperforms state-of-the-art by at least 6.3%. For low complexity calibration sets, our method outperforms other calibration methods. More complex calibration algorithms do not outperform our method until the size of the calibration set is excessively large. Even then, the gains obtained by alternatives are small, e.g., only 0.1 degree lower error for 64 gaze targets. Source code is available at https://github.com/czk32611/Gaze-Decomposition.

****

Detecting Morphed Face Attacks Using Residual Noise from Deep Multi-Scale Context Aggregation Network

Sushma Venkatesh, Raghavendra Ramachandra, Kiran Raja, Luuk Spreeuwers, Raymond Veldhuis, Christoph Busch; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 280-289

The evolving deployment of face recognition system has raised concerns regarding the vulnerability of those systems to various attacks. The morphed face attack involves two different face images via morphing to obtain an attack image face image similar to both original images. The obtained morphed image can easily be verified against both subjects visually and by Face Recognition Systems (FRS). The face morphing attack thus raises a severe concern to various security applications like border control and e-passport unless such attacks are detected and mitigated. In this work, we propose a novel method to reliably detect the morphed face attacks using a new denoising framework. Considering the time complexity and parameterization efforts, we realize the proposed method using a new deep Multi-scale Context Aggregation Network (MS-CAN). Extensive experiments are carried out on three different morphed face image datasets. The Morphed Attack Detection (MAD) performance of the proposed method is also benchmarked against 13 different state-of-the-art techniques using the ISO IEC 30107-3 evaluation metrics. Based on the obtained quantitative results reported, the proposed method has indicated the best performance.

****

Gaze Estimation for Assisted Living Environments

Philipe Ambrozio Dias, Damiano Malafronte, Henry Medeiros, Francesca Odone; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 290-299

Effective assisted living environments must be able to perform inferences on how their occupants interact with one another as well as with surrounding objects. To accomplish this goal using a vision-based automated approach, multiple tasks such as pose estimation, object segmentation and gaze estimation must be addressed. Gaze direction provides some of the strongest indications of how a person interacts with the environment. In this paper, we propose a simple neural network regressor that estimates the gaze direction of individuals in a multi-camera assisted living scenario, relying only on the relative positions of facial keypoints collected from a single pose estimation model. To handle cases of keypoint occlusion, our model exploits a novel confidence gated unit in its input layer. In addition to the gaze direction, our model also outputs an estimation of its own prediction uncertainty. Experimental results on a public benchmark demonstrate that our approach performs on par with a complex, dataset-specific baseline, while its uncertainty predictions are highly correlated to the actual angular error of corresponding estimations. Finally, experiments on images from a real assisted living environment demonstrate that our model has a higher suitability for its final application.

****

On Hallucinating Context and Background Pixels from a Face Mask using Multi-scale GANs

Sandipan Banerjee, Walter Scheirer, Kevin Bowyer, Patrick Flynn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 300-309

We propose a multi-scale GAN model to hallucinate realistic context (forehead, hair, neck, clothes) and background pixels automatically from a single input face mask, without any user supervision. Instead of swapping a face on to an existing picture, our model directly generates realistic context and background pixels based on the features of the provided face mask. Unlike facial inpainting algorithms, it can generate realistic hallucinations even for a large number of missing pixels. Our model is composed of a cascaded network of GAN blocks, each tasked with hallucination of missing pixels at a particular resolution while guiding the synthesis process of the next GAN block. The hallucinated full face image is made photo realistic by using a combination of reconstruction, perceptual, adversarial and identity preserving losses at each block of the network. With a set of extensive experiments, we demonstrate the effectiveness of our model in hallucinating context and background pixels from face masks varying in facial pose, expression and lighting, collected from multiple datasets subject disjoint with our training data. We also compare our method with popular face inpainting and face swapping models in terms of visual quality, realism and identity preservation. Additionally, we analyze our cascaded pipeline and compare it with the progressive growing of GANs, and explore its usage as a data augmentation module for training CNNs.

****

EyeGAN: Gaze-Preserving, Mask-Mediated Eye Image Synthesis

Harsimran Kaur, Roberto Manduchi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 310-319

Automatic synthesis of realistic eye images with prescribed gaze direction is important for multiple application domains. We introduce EyeGAN, an algorithm to generate eye images in the style of a desired target domain, that inherit annotations available in images from a source domain. EyeGAN takes in input ternary masks, which are used as domain-independent proxies for gaze direction. We evaluate EyeGAN against competing eye image synthesis algorithms by measuring a specific gaze consistency index. In addition, we present results from multiple experiments (involving eye region segmentation, pupil localization, and gaze direction estimation) showing that the use of EyeGAN generated images with inherited annotations for network training leads to superior performances compared to other domain transfer algorithms.

****

Boosting Deep Face Recognition via Disentangling Appearance and Geometry

Ali Dabouei, Fariborz Taherkhani, Sobhan Soleymani, Jeremy Dawson, Nasser Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 320-329

In this paper, we propose a framework for disentangling the appearance and geometry representations in the face recognition task. To provide supervision for this aim, we generate geometrically identical faces by incorporating spatial transformations. We demonstrate that the proposed approach enhances the performance of deep face recognition models by assisting the training process in two ways. First, it enforces the early and intermediate convolutional layers to learn more representative features that satisfy the properties of disentangled embeddings. Second, it augments the training set by altering faces geometrically. Through extensive experiments, we demonstrate that integrating the proposed approach into state-of-the-art face recognition methods effectively improves their performance on challenging datasets, such as LFW, YTF, and MegaFace. Both theoretical and practical aspects of the method are analyzed rigorously by concerning ablation studies and knowledge transfer tasks. Furthermore, we show that the knowledge leaned by the proposed method can favor other face-related tasks, such as attribute prediction.

****

Robust Facial Landmark Detection via Aggregation on Geometrically Manipulated Faces
Seyed Mehdi Iranmanesh, Ali Dabouei, Sobhan Soleymani, Hadi Kazemi, Nasser Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 330-340
In this work, we present a practical approach to the problem of facial landmark detection. The proposed method can deal with large shape and appearance variations under the rich shape deformation. To handle the shape variations we equip our method with the aggregation of manipulated face images. The proposed framework generates different manipulated faces using only one given face image. The approach utilizes the fact that small but carefully crafted geometric manipulation in the input domain can fool deep face recognition models.   We propose three different approaches to generate manipulated faces in which two of them perform the manipulations via adversarial attacks and the other one uses known transformations. Aggregating the manipulated faces provides a more robust landmark detection approach which is able to capture more important deformations and variations of the face shapes. Our approach is demonstrated its superiority compared to the state-of-the-art method on benchmark datasets AFLW, 300-W, and COFW.
****
End to End Lip Synchronization with a Temporal AutoEncoder
Yoav Shalev, Lior Wolf; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 341-350
We study the problem of syncing the lip movement in a video with the audio stream. Our solution finds an optimal alignment using a dual-domain recurrent neural network that is trained on synthetic data we generate by dropping and duplicating video frames. Once the alignment is found, we modify the video in order to sync the two sources. Our method is shown to greatly outperform the literature methods on a variety of existing and new benchmarks. As an application, we demonstrate our ability to robustly align text-to-speech generated audio with an existing video stream. Our code is attached as supplementary.
****
Can a CNN Automatically Learn the Significance of Minutiae Points for Fingerprint Matching?
Anurag Chowdhury, Simon Kirchgasser, Andreas Uhl, Arun Ross; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 351-359
Most automated fingerprint recognition systems use minutiae points for comparing fingerprints. In the parlance of Computer Vision, minutiae can be viewed as handcrafted features, i.e., features that have been proposed by human experts for the task of fingerprint recognition. In this work, we raise the following question: Can a machine learning system automatically determine the significance of minutiae points for fingerprint matching? To this effect, a patch-based Siamese Convolutional Neural Network (CNN), which does not explicitly rely on the extraction of minutiae points, is designed and trained from scratch. The purpose of this network is to learn the most effective features for matching fingerprint images. The features learned by this network are analyzed using Gradient-weighted Class Activation Mapping (Grad-CAM) to determine if they correlate with the locations of minutiae points. Our experiments suggest that the proposed network automatically learns to focus on minutiae points, when available, for fingerprint matching. Thus, an automated learner without any explicit domain knowledge establishes the significance of minutiae points for fingerprint matching.
****
AutoToon: Automatic Geometric Warping for Face Cartoon Generation
Julia Gong, Yannick Hold-Geoffroy, Jingwan Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 360-369
Caricature, a type of exaggerated artistic portrait, amplifies the distinctive, yet nuanced traits of human faces. This task is typically left to artists, as it has proven difficult to capture subjects' unique characteristics well using automated methods. Recent development of deep end-to-end methods has achieved promising results in capturing style and higher-level exaggerations. However, a key p

art of caricatures, face warping, has remained challenging for these systems. In this work, we propose AutoToon, the first supervised deep learning method that yields high-quality warps for the warping component of caricatures. Completely disentangled from style, it can be paired with any stylization method to create diverse caricatures. In contrast to prior art, we leverage an SENet and spatial transformer module and train directly on artist warping fields, applying losses both prior to and after warping. As shown by our user studies, we achieve appealing exaggerations that amplify distinguishing features of the face while preserving facial detail.
****

Component Attention Guided Face Super-Resolution Network: CAGFace
Ratheesh Kalarot, Tao Li, Fatih Porikli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 370-380
To make the best use of the underlying structure of faces, the collective information through face datasets and the intermediate estimates during the upsampling process, here we introduce a fully convolutional multi-stage neural network for 4x super-resolution for face images. We implicitly impose facial component-wise attention maps using a segmentation network to allow our network to focus on face-inherent patterns. Each stage of our network is composed of a stem layer, a residual backbone, and spatial upsampling layers. We recurrently apply stages to reconstruct an intermediate image, and then reuse its space-to-depth converted versions to bootstrap and enhance image quality progressively. Our experiments show that our face super-resolution method achieves quantitatively superior and perceptually pleasing results in comparison to state of the art.
****

Nonparametric Structure Regularization Machine for 2D Hand Pose Estimation
Yifei Chen, Haoyu Ma, Deying Kong, Xiangyi Yan, Jianbao Wu, Wei Fan, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 381-390
Hand pose estimation is more challenging than body pose estimation due to severe articulation, self-occlusion and high dexterity of the hand. Current approaches often rely on a popular body pose algorithm, such as the Convolutional Pose Machine (CPM), to learn 2D keypoint features. These algorithms cannot adequately address the unique challenges of hand pose estimation, because they are trained solely based on keypoint positions without seeking to explicitly model structural relationship between them. We propose a novel Nonparametric Structure Regularization Machine (NSRM) for 2D hand pose estimation, adopting a cascade multi-task architecture to learn hand structure and keypoint representations jointly. The structure learning is guided by synthetic hand mask representations, which are directly computed from keypoint positions, and is further strengthened by a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy of mask synthesis. We conduct extensive studies on two public datasets - OneHand 10k and CMU Panoptic Hand. Experimental results demonstrate that explicitly enforcing structure learning consistently improves pose estimation accuracy of CPM baseline models, by 1.17% on the first dataset and 4.01% on the second one. The implementation and experiment code is freely available online. Our proposal of incorporating structural learning to hand pose estimation requires no additional training information, and can be a generic add-on module to other pose estimation models.
****

3D Hand Pose Estimation with Disentangled Cross-Modal Latent Space
Jiajun Gu, Zhiyong Wang, Wanli Ouyang, weichen zhang, Jiafeng Li, Li Zhuo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 391-400
Estimating 3D hand pose from a single RGB image is a challenging task because of its ill-posed nature (i.e., depth ambiguity). Recently, various generative-based approaches have been proposed to predict the 3D joints by learning a unified latent space between two modalities (i.e., RGB image and 3D joints). However, projecting multi-modal data(i.e., RGB images and 3D joints) into a unified latent space is difficult as the modality-specific features usually inter-fere t

he learning of the optimal latent space. Hence in this paper, we propose to disentangle the latent space into two sub-latent spaces: modality-specific latent space and pose-specific latent space for 3D hand pose estimation. Our proposed method, namely Disentangled Cross-Modal Latent Space (DCMLS), consists of two variational autoencoder networks and auxiliary components which connects the two VAEs to align underlying hand poses and transfer modality context from RGB to 3D. For the hand pose latent space,we align the hand pose latent space from the two modalities by using a cross-modal discriminator with the adversarial learning strategy. For the context latent space, we learn acontext translator to gain access to the cross-modal con-text. Experimental results on two widely used public bench-mark datasets RHD and STB demonstrate that our proposed DCMLS method is able to outperform the state-of-the-artones on single image based 3D hand pose estimation.

****

Robust Template-Based Non-Rigid Motion Tracking Using Local Coordinate Regularization

Wei Li, Shang Zhao, Xiao Xiao, James Hahn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 401-410

In this paper, we propose our template-based non-rigid registration algorithm to address the misalignments in the frame-to-frame motion tracking with single or multiple commodity depth cameras. We analyze the deformation in the local coordinates of neighboring nodes and use this differential representation to formulate the regularization term for the deformation field in our non-rigid registration. The local coordinate regularizations vary for each pair of neighboring nodes based on the tracking status of the surface regions. We propose our tracking strategies for different surface regions to minimize misalignments and reduce error accumulation. This method can thus preserve local geometric features and prevent undesirable distortions. Moreover, we introduce a geodesic-based correspondence estimation algorithm to align surfaces with large displacements. Finally, we demonstrate the effectiveness of our proposed method with detailed experiments.

****

DGGAN: Depth-image Guided Generative Adversarial Networks for Disentangling RGB and Depth Images in 3D Hand Pose Estimation

Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, Wei Fan, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 411-419

Estimating3D hand poses from RGB images is essentialto a wide range of potential applications, but is challengingowing to substantial ambiguity in the inference of depth in-formation from RGB images. State-of-the-art estimators ad-dress this problem by regularizing3D hand pose estimationmodels during training to enforce the consistency betweenthe predicted3D poses and the ground truth depth maps.However, these estimators rely on the availability of bothRGB images and paired depth maps during training. In thisstudy, we propose a conditional generative adversarial net-work model, called Depth-image Guided GAN (DGGAN),to generate realistic depth maps conditioned on the inputRGB image, and use the synthesized depth maps to regular-ize the3D hand pose estimation model, therefore eliminat-ing the need for ground truth depth maps. Experimental re-sults on multiple benchmark datasets show that the synthe-sized depth maps produced by DGGAN are quite effective inregularizing the pose estimation model, yielding new state-of-the-art results in estimation accuracy, notably reducingthe mean3D end-point errors (EPE) by4.7%,16.5%, and6.8%on the RHD, STB and MHP datasets, respectively.

****

Multiview Supervision By Registration

Yilun Zhang, Hyun Soo Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 420-428

This paper presents a semi-supervised learning framework to train a keypoint detector using multiview image streams given the limited labeled data (typically <4%). We leverage the complementary relationship between multiview geometry and visual tracking to provide three types of supervisionary signals to utilize the unlabeled data: (1) keypoint detection in one view can be supervised by other view s via the epipolar geometry; (2) a keypoint moves smoothly over time where its optical flow can be used to temporally supervise consecutive image frames to each other; (3) visible keypoint in one view is likely to be visible in the adjacent view. We integrate these three signals in a differentiable fashion to design a new end-to-end neural network composed of three pathways. This design allows us to extensively use the unlabeled data to train the keypoint detector. We show that our approach outperforms existing detectors including DeepLabCut tailored to the keypoint detection of non-human species such as monkeys, dogs, and mice.

****

DeepFuse: An IMU-Aware Network for Real-Time 3D Human Pose Estimation from Multi-View Image

Fuyang Huang, Ailing Zeng, Minhao Liu, Qiuxia Lai, Qiang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 429-438

In this paper, we propose a two-stage fully 3D network, namely DeepFuse, to estimate human pose in 3D space by fusing body-worn Inertial Measurement Unit (IMU) data and multi-view images deeply. The first stage is designed for pure vision estimation. To preserve data primitiveness of multi-view inputs, the vision stage uses multi-channel volume as data representation and 3D soft-argmax as activation layer. The second one is the IMU refinement stage which introduces an IMU-bone layer to fuse the IMU and vision data earlier at data level. without requiring a given skeleton model a priori, we can achieve a mean joint error of 28.9mm on TotalCapture dataset and 13.4mm on Human3.6M dataset under protocol 1, improving the SOTA result by a large margin. Finally, we discuss the effectiveness of a fully 3D network for 3D pose estimation experimentally which may benefit future research.

****

Attention-based Fusion for Multi-source Human Image Generation

Stephane Lathuiliere, Enver Sangineto, Aliaksandr Siarohin, Nicu Sebe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 439-448

We present a generalization of the person-image generation task, in which a human image is generated conditioned on a target pose and a set X of source appearance images. In this way, we can exploit multiple, possibly complementary images of the same person which are usually available at training and at testing time. The solution we propose is mainly based on a local attention mechanism which selects relevant information from different source image regions, avoiding the necessity to build specific generators for each specific cardinality of X. The empirical evaluation of our method shows the practical interest of addressing the person-image generation problem in a multi-source setting.

****

Real-Time Multi-Person Pose Tracking using Data Assimilation

Caterina Buizza, Tobias Fischer, Yiannis Demiris; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 449-458

We propose a framework for the integration of data assimilation and machine learning methods in human pose estimation, with the aim of enabling any pose estimation method to be run in real-time, whilst also increasing consistency and accuracy. Data assimilation and machine learning are complementary methods: the former allows us to make use of information about the underlying dynamics of a system but lacks the flexibility of a data-based model, which we can instead obtain with the latter. Our framework presents a real-time tracking module for any single or multi-person pose estimation system. Specifically, tracking is performed by a number of Kalman filters initiated for each new person appearing in a motion sequence. This permits tracking of multiple skeletons and reduces the frequency that computationally expensive pose estimation has to be run, enabling online pose tracking. The module tracks for N frames while the pose estimates are calculated for frame (N+1). This also results in increased consistency of person identification and reduced inaccuracies due to missing joint locations and inversion of left-and right-side joints.

****

Reducing Footskate in Human Motion Reconstruction with Ground Contact Constraints
Yuliang Zou, Jimei Yang, Duygu Ceylan, Jianming Zhang, Federico Perazzi, Jia-Bin Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 459-468
In this paper, we aim to reduce the footskate artifacts when reconstructing human dynamics from monocular RGB videos. Recent work has made substantial progress in improving the temporal smoothness of the reconstructed motion trajectories. Their results, however, still suffer from severe foot skating and slippage artifacts. To tackle this issue, we present a neural network based detector for localizing ground contact events of human feet and use it to impose a physical constraint for optimization of the whole human dynamics in a video. We present a detailed study on the proposed ground contact detector and demonstrate high-quality human motion reconstruction results in various videos.
****
Unsupervised Cross-Dataset Adaptation via Probabilistic Amodal 3D Human Pose Completion
Jogendra Nath Kundu, Rahul M V, Jay Patravali, Venkatesh Babu RADHAKRISHNAN; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 469-478
Despite remarkable success of supervised deep learning models for 3D human pose estimation, performance of such models is mostly limited to constrained laboratory settings. Such models not only exhibit an alarming level of dataset bias but also fail to operate on unconstrained videos in the presence of external variations such as camera motion, partial body visibility, occlusion, etc. Acknowledging these shortcomings, firstly, we aim to formalize a motion representation learning framework by effectively utilizing both constrained and artificially generated unconstrained video samples for datasets with 3D pose annotation. Without ignoring the inherent uncertainty in pose estimation for the truncated video frames, we devise a novel probabilistic amodal pose completion framework to enable generation of multiple plausible pose-filling outcomes. Secondly, to address dataset bias, the probabilistic amodal framework is re-utilized to design novel self-supervised objectives. This not only enables adaptation of the model to target unannotated datasets (wild YouTube videos) but also encourages learning of generic motion representations beyond the available supervised data even in unconstrained scenarios. Such a training regime helps us achieve state-of-the-art performance on unsupervised cross-dataset pose estimation, with a significant improvement in partially-visible unconstrained scenarios.
****
Lightweight 3D Human Pose Estimation Network Training Using Teacher-Student Learning
Dong-Hyun Hwang, Suntae Kim, Nicolas Monet, Hideki Koike, Soonmin Bae; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 479-488
We present MoVNect, a lightweight deep neural network to capture 3D human pose using a single RGB camera. To improve the overall performance of the model, we apply the teacher-student learning method based knowledge distillation to 3D human pose estimation. Real-time post-processing makes the CNN output yield temporally stable 3D skeletal information, which can be used in applications directly. We implement a 3D avatar application running on mobile in real-time to demonstrate that our network achieves both high accuracy and fast inference time. Extensive evaluations show the advantages of our lightweight model with the proposed training method over previous 3D pose estimation methods on the Human3.6M dataset and mobile devices.
****
Detecting the Starting Frame of Actions in Video
Iljung Kwak, Jian-Zhong Guo, Adam Hantman, David Kriegman, Kristin Branson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 489-497
In this work, we address the problem of precisely localizing key frames of an ac

tion, for example, the precise time that a pitcher releases a baseball, or the precise time that a crowd begins to applaud. Key frame localization is a largely overlooked and important action-recognition problem, for example in the field of neuroscience, in which we would like to understand the neural activity that produces the start of a bout of an action. To address this problem, we introduce a novel structured loss function that properly weights the types of errors that matter in such applications: it more heavily penalizes extra and missed action start detections over small misalignments. Our structured loss is based on the best matching between predicted and labeled action starts. We train recurrent neural networks (RNNs) to minimize differentiable approximations of this loss. To evaluate these methods, we introduce the Mouse Reach Dataset, a large, annotated video dataset of mice performing a sequence of actions. The dataset was collected and labeled by experts for the purpose of neuroscience research. On this dataset, we demonstrate that our method outperforms related approaches and baseline methods using an unstructured loss.

****

Looking deeper into Time for Activities of Daily Living Recognition

Srijan Das, Monique Thonnat, Francois Bremond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 498-507

In this paper, we introduce a new approach for Activities of Daily Living (ADL) recognition. In order to discriminate between activities with similar appearance and motion, we focus on their temporal structure. Actions with subtle and similar motion are hard to disambiguate since long-range temporal information is hard to encode. So, we propose an end-to-end Temporal Model to incorporate long-range temporal information without losing subtle details. The temporal structure is represented globally by different temporal granularities and locally by temporal segments. We also propose a two-level pose driven attention mechanism to take into account the relative importance of the segments and granularities. We validate our approach on 2 public datasets: a 3D human activity dataset (NTU-RGB+D) and a human-object interaction dataset (Northwestern-UCLA Multiview Action 3D). Our Temporal Model can also be incorporated with any existing 3D CNN (including attention based) as a backbone which reveals its robustness.

****

Weakly-Supervised Multi-Person Action Recognition in 360° Videos

Junnan Li, Jianquan Liu, Wong Yongkang, Shoji Nishimura, Mohan Kankanhalli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 508-516

The recent development of commodity 360^ \circ cameras have enabled a single video to capture an entire scene, which endows promising potentials in surveillance scenarios. However, research in omnidirectional video analysis has lagged behind the hardware advances. In this work, we address the important problem of action recognition in top-view 360^ \circ videos. Due to the wide filed-of-view, 360^ \circ videos usually capture multiple people performing actions at the same time. Furthermore, the appearance of people are deformed. The proposed framework first transforms omnidirectional videos into panoramic videos, then it extracts spatial-temporal features using region-based 3D CNNs for action recognition. We propose a weakly-supervised method based on multi-instance multi-label learning, which trains the model to recognize and localize multiple actions in a video using only video-level action labels as supervision. We perform experiments to quantitatively validate the efficacy of the proposed method and qualitatively demonstrate action localization results. To enable research in this direction, we introduce 360Action, the first omnidirectional video dataset for multi-person action recognition.

****

Learning Multimodal Representations for Unseen Activities

AJ Piergiovanni, Michael Ryoo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 517-526

We present a method to learn a joint multimodal representation space that enables recognition of unseen activities in videos. We first compare the effect of placing various constraints on the embedding space using paired text and video data

. We also propose a method to improve the joint embedding space using an adversarial formulation, allowing it to benefit from unpaired text and video data. By using unpaired text data, we show the ability to learn a representation that better captures unseen activities. In addition to testing on publicly available datasets, we introduce a new, large-scale text/video dataset. We experimentally confirm that using paired and unpaired data to learn a shared embedding space benefits three difficult tasks (i) zero-shot activity classification, (ii) unsupervised activity discovery, and (iii) unseen activity captioning, outperforming the state-of-the-arts.

****

Actor Conditioned Attention Maps for Video Action Detection

Oytun Ulutan, Swati Rallapalli, Mudhakar Srivatsa, Carlos Torres, B. S. Manjunath; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 527-536

While observing complex events with multiple actors, humans do not assess each actor separately, but infer from the context. The surrounding context provides essential information for understanding actions. To this end, we propose to replace region of interest(RoI) pooling with an attention module, which ranks each spatio-temporal region's relevance to a detected actor instead of cropping. We refer to these as Actor-Conditioned Attention Maps (ACAM), which amplify/dampen the features extracted from the entire scene. The resulting actor-conditioned features focus the model on regions that are relevant to the conditioned actor. For actor localization, we leverage pre-trained object detectors, which transfer better. The proposed model is efficient and our action detection pipeline achieves near real-time performance. Experimental results on AVA 2.1 and JHMDB demonstrate the effectiveness of attention maps, with improvements of 7 mAP on AVA and 4 mAP on JHMDB.

****

Weakly Supervised Gaussian Networks for Action Detection

Basura Fernando, Cheston Tan, Hakan Bilen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 537-546

Detecting temporal extents of human actions in videos is a challenging computer vision problem that requires detailed manual supervision including frame-level labels. This expensive annotation process limits deploying action detectors to a limited number of categories. We propose a novel method, called WSGN, that learns to detect actions from weak supervision, using only video-level labels. WSGN learns to exploit both video-specific and dataset-wide statistics to predict relevance of each frame to an action category. This strategy leads to significant gains in action detection for two standard benchmarks THUMOS14 and Charades. Our method obtains excellent results compared to state-of-the-art methods that uses similar features and loss functions on THUMOS14 dataset. Similarly, our weakly supervised method is only 0.3% mAP behind a state-of-the-art supervised method on challenging Charades dataset for action localization.

****

Weakly Supervised Temporal Action Localization Using Deep Metric Learning

Ashraful Islam, Richard Radke; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 547-556

Temporal action localization is an important step towards video understanding. Most current action localization methods depend on untrimmed videos with full temporal annotations of action instances. However, it is expensive and time-consuming to annotate both action labels and temporal boundaries of videos. To this end, we propose a weakly supervised temporal action localization method that only requires video-level action instances as supervision during training. We propose a classification module to generate action labels for each segment in the video, and a deep metric learning module to learn the similarity between different action instances. We jointly optimize a balanced binary cross-entropy loss and a metric loss using a standard backpropagation algorithm. Extensive experiments demonstrate the effectiveness of both of these components in temporal localization. We evaluate our algorithm on two challenging untrimmed video datasets: THUMOS14 and ActivityNet1.2. Our approach improves the current state-of-the-art result fo r THUMOS14 by 6.5% mAP at IoU threshold 0.5, and achieves competitive performance for ActivityNet1.2.

****

Dynamic Motion Representation for Human Action Recognition

Sadjad Asghari-Esfeden, Mario Sznaier, Octavia Camps; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 557-566

Despite the advances in Human Activity Recognition, the ability to exploit the dynamics of human body motion in videos has yet to be achieved. In numerous recent works, researchers have used appearance and motion as independent inputs to infer the action that is taking place in a specific video. In this paper, we highlight that while using a novel representation of human body motion, we can benefit from appearance and motion simultaneously. As a result, better performance of action recognition can be achieved. We start with a pose estimator to extract the location and heat-map of body joints in each frame. We use a dynamic encoder to generate a fixed size representation from these body joint heat-maps. Our experimental results show that training a convolutional neural network with the dynamic motion representation outperforms state-of-the-art action recognition models. By modeling distinguishable activities as distinct dynamical systems and with the help of two stream networks, we obtain the best performance on HMDB, JHMDB, UCF-101, and AVA datasets.

****

Image to Video Domain Adaptation Using Web Supervision

Andrew Kae, Yale Song; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 567-575

Training deep neural networks typically requires large amounts of labeled data which may be scarce or expensive to obtain for a particular target domain. As an alternative, we can leverage webly-supervised data (i.e. results from a public search engine) which are relatively plentiful but may contain noisy results. In this work, we propose a novel two-stage approach to learn a video classifier using webly-supervised data. We argue that learning appearance features and temporal features sequentially, rather than jointly, is an easier optimization for this task. We show this by first learning an image model from web images, which is used to initialize and train a video model. Our model applies domain adaptation to account for potential domain shift present between the source domain (webly-supervised data) and target domain, and also accounts for noise by adding a novel attention component. We report results competitive with state-of-the-art for webly-supervised approaches (while simplifying the training process) on UCF-101 and also evaluate on Kinetics for comparison.

****

Stacked Spatio-Temporal Graph Convolutional Networks for Action Segmentation

Pallabi Ghosh, Yi Yao, Larry Davis, Ajay Divakaran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 576-585

We propose novel Stacked Spatio-Temporal Graph Convolutional Networks (Stacked-STGCN) for action segmentation, i.e., predicting and localizing a sequence of actions over long videos. We extend the Spatio-Temporal Graph Convolutional Network (STGCN) originally proposed for skeleton-based action recognition to enable nodes with different characteristics (e.g., scene, actor, object, action), feature descriptors with varied lengths, and arbitrary temporal edge connections to account for large graph deformation commonly associated with complex activities. We further introduce the stacked hourglass architecture to STGCN to leverage the advantages of an encoder-decoder design for improved generalization performance and localization accuracy. We explore various descriptors such as frame-level VGG, segment-level I3D, RCNN-based object, etc. as node descriptors to enable action segmentation based on joint inference over comprehensive contextual information. We show results on CAD120 (which provides pre-computed node features and edge weights for fair performance comparison across algorithms) as well as a more complex real world activity dataset, Charades. Our Stacked-STGCN in general achieves improved performance over the state-of-the-art for both CAD120 and Charades. M oreover, due to its generic design, Stacked-STGCN can be applied to a wider range of applications that require structured inference over long sequences with heterogeneous data types and varied temporal extent.

****

Global Co-occurrence Feature Learning and Active Coordinate System Conversion for Skeleton-based Action Recognition

Sheng Li, Tingting Jiang, Tiejun Huang, Yonghong Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 586-594

Skeleton-based action recognition has attracted more and more attention in recent years. Besides, the rapid development of deep learning has greatly improved the performance.However, the current exploration of action cooccurrence is still not comprehensive enough. Most existing works only mine co-occurrence features from the temporal or spatial domain seperately, and it's common to combine them in the end. Different from previous works, our approach is able to learn temporal and spatial co-occurrence features integratedly and globally, which is called spatio-temporal-unit feature enhancement (STUFE). In order to better align the skeleton data, we introduce a novel method for skeleton data preprocessing called active coordinate system conversion (ACSC). A coordinate system can be learned automatically to transform skeleton samples for alignment. By the way, the proposed methods are compatible with current two types of mainstream models, the CNN-based and GCN-based models. Finally, on the two benchmarks of NTU-RGB+D and SBU Kinect Interaction, we validated our methods based on two mainstream models.The results show that our methods achieve the state-of-the-art.

****

Few-Shot Learning of Video Action Recognition Only Based on Video Contents

Yang Bo, Yangdi Lu, Wenbo He; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 595-604

The success of video action recognition based on Deep Neural Networks (DNNs) is highly dependent on a large number of manually labeled videos. In this paper, we introduce a supervised learning approach to recognize video actions with very few training videos. Specifically, we propose Temporal Attention Vectors (TAVs) which adapt various length videos to preserve the temporal information of the entire video. We evaluate the TAVs on UCF101 and HMDB51. Without training any deep 3D or 2D frame feature extractors on video datasets (only pre-trained on ImageNet), the TAVs only introduce 2.1M parameters but outperforms the state-of-the-art video action recognition benchmarks with very few labeled training videos (e.g. 92% on UCF101 and 59% on HMDB51, with 10 and 8 training videos per class, respectively). Furthermore, our approach can still achieve competitive results on full datasets (97.1% on UCF101 and 77% on HMDB51).

****

Action Segmentation with Mixed Temporal Domain Adaptation

Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 605-614

The main progress for action segmentation comes from densely-annotated data for fully-supervised learning. Since manual annotation for frame-level actions is timeconsuming and challenging, we propose to exploit auxiliary unlabeled videos, which are much easier to obtain, by shaping this problem as a domain adaptation (DA) problem. Although various DA techniques have been proposed in recent years, most of them have been developed only for the spatial direction. Therefore, we propose Mixed Temporal Domain Adaptation (MTDA) to jointly align frame- and video-level embedded feature spaces across domains, and further integrate with the domain attention mechanism to focus on aligning the frame-level features with higher domain discrepancy, leading to more effective domain adaptation. Finally, we evaluate our proposed methods on three challenging datasets (GTEA, 50Salads, and Breakfast), and validate that MTDA outperforms the current state-of-the-art methods on all three datasets by large margins (e.g. 6.4% gain on F1@50 and 6.8% gain on the edit score for GTEA).

****

Action Graphs: Weakly-supervised Action Localization with Graph Convolution Networks

Maheen Rashid, Hedvig Kjellstrom, Yong Jae Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 615-624

We present a method for weakly-supervised action localization based on graph convolutions. In order to find and classify video time segments that correspond to relevant action classes, a system must be able to both identify discriminative time segments in each video, and identify the full extent of each action. Achieving this with weak video level labels requires the system to use similarity and dissimilarity between moments across videos in the training data to understand both how an action appears, as well as the sub-actions that comprise the action's full extent. However, current methods do not make explicit use of similarity between video moments to inform the localization and classification predictions. We present a novel method that uses graph convolutions to explicitly model similarity between video moments. Our method utilizes similarity graphs that encode appearance and motion, and pushes the state of the art on THUMOS `14, ActivityNet 1.2, and Charades for weakly-supervised action localization.

****

D3D: Distilled 3D Networks for Video Action Recognition

Jonathan Stroud, David Ross, Chen Sun, Jia Deng, Rahul Sukthankar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 625-634

State-of-the-art methods for action recognition commonly use two networks: the spatial stream, which takes RGB frames as input, and the temporal stream, which takes optical flow as input. In recent work, both streams are 3D Convolutional Neural Networks, which extract features using spatiotemporal filters. These filters can respond to motion, and therefore should allow the network to learn motion representations, removing the need for optical flow. However, we still see significant benefits in performance by feeding optical flow into the temporal stream, indicating that the spatial stream is "missing" some of the signal that the temporal stream captures. In this work, we first investigate whether motion representations are indeed missing in the spatial stream, and show that there is significant room for improvement. Second, we demonstrate that these motion representations can be improved using distillation, that is, by tuning the spatial stream to mimic the temporal stream, effectively combining both models into a single stream. Finally, we show that our Distilled 3D Network (D3D) achieves performance on par with the two-stream approach, with no need to compute optical flow during inference.

****

Self-Attention Network for Skeleton-based Human Action Recognition

Sangwoo Cho, Muhammad Maqbool, Fei Liu, Hassan Foroosh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 635-644

Skeleton-based action recognition has recently attracted a lot of attention. Researchers are coming up with new approaches for extracting spatio-temporal relations and making considerable progress on large-scale skeleton based datasets. Most of the architectures being proposed are based upon recurrent neural networks (RNNs), convolutional neural networks (CNNs) and graph-based CNNs. When it comes to skeleton-based action recognition, the importance of long term contextual information is central which is not captured by the current architectures. In order to come up with a better representation and capturing of long term spatio-temporal relationships, we propose three variants of Self-Attention Network (SAN), namely, SAN-V1,SAN-V2 and SAN-V3. Our SAN variants has the impressive capability of extracting high-level semantics by capturing long-range correlations. We have also integrated the Temporal Segment Network (TSN) with our SAN variants which resulted in improved overall performance. Different configurations of Self-Attention Network (SAN) variants and Temporal Segment Network (TSN) are explored with extensive experiments. Our chosen configuration outperforms state-of-the-art Top-1 and Top-5 by 4.4% and 2.9% respectively on Kinetics and shows consistently better performance than state-of-the-art methods on N

TU RGB+D.

****

Long-Short Graph Memory Network for Skeleton-based Action Recognition

Junqin Huang, zhenhuan huang, Xiang Xiang, Xuan Gong, Baochang Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 645-652

Current studies have shown the effectiveness of long short-term memory network (LSTM) for skeleton-based human action recognition in capturing temporal and spatial features of the skeleton sequence. Nevertheless, it still remains challenging for LSTM to extract the latent structural dependency among nodes. In this paper, we introduce a new long-short graph memory network (LSGM) to improve the capability of LSTM to model the skeleton sequence - a type of graph data. Our proposed LSGM can learn high-level temporal-spatial features end-to-end, enabling LSTM to extract the spatial information that is neglected but intrinsic to the skeleton graph data. To improve the discriminative ability of the temporal and spatial module, we use a calibration module termed as graph temporal-spatial calibration (GTSC) to calibrate the learned temporal-spatial features. By integrating the two modules into the same framework, we obtain a stronger generalization capability in processing dynamic graph data and achieve a significant performance improvement on the NTU and SYSU dataset. Experimental results have validated the effectiveness of our proposed LSGM+GTSC model in extracting temporal and spatial information from dynamic graph data.

****

Weakly Supervised Graph Convolutional Neural Network for Human Action Localization

Daisuke Miki, Shi Chen, Kazuyuki Demachi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 653-661

Skeleton-based human action recognition from video sequences is currently an active topic of research. Conventionally, human action recognition is performed after conducting feature extraction on a given spatial-temporal representation of a human pose by using statistical methods or deep learning methods. The spatial and temporal features are globally evaluated by a classifier and used to determine which action is closest. However, the conventional methodology does not identify the temporal location of the action that determines the classification. To address this problem, we propose a skeleton-based human action recognition and localization method using weakly supervised graph convolutional neural networks, which are both spatially and temporally connected. In this method, human action localization is accomplished using time series data of human joint positions as input and then applying regression to find an expected value for each action at each time frame. Our weakly supervised training is based on multiple-instance learning inspired by deep ranking, and we devise a loss function so that high scores can be spontaneously learned for temporally important time frames. In this paper, we first explain the network architecture and then present a multiple-instance learning method for its optimization. In the experiment, we performed localization and classification of human actions by using this method and confirmed the temporal localization efficacy of the method.

****

Temporal Contrastive Pretraining for Video Action Recognition

Guillaume LORRE, Jaonary Rabarisoa, Astrid Orcesi, Samia Ainouz, Stephane Canu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 662-670

In this paper, we propose a self-supervised method for video representation learning based on Contrastive Predictive Coding (CPC) [27]. Previously, CPC has been used to learn representations for different signals (audio, text or image). It benefits from the use of an autoregressive modeling and contrastive estimation to learn long-term relations inside raw signal while remaining robust to local noise. Our self-supervised task consists in predicting the latent representation of future segments of the video. As opposed to generative models, predicting directly in the feature space is easier and avoid incertitude problems for long-term predictions. Today, using CPC to learn representations for videos remains chall enging due to the structure and the high dimensionality of the signal. We demonstrate experimentally that the representations learned by the network are useful for action recognition. We test it with different input types such as optical flows, image differences and raw images on different datasets (UCF-101 and HMDB51). It gives consistent results across the modalities. At last, we notice the utility of our pre-training method by achieving competitive results for action recognition using few labeled data.

****

Fine-Grained Motion Representation For Template-Free Visual Tracking
Kai Shuang, Yuheng Huang, Yue Sun, Zhun Cai, Hao Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 671-680

The object tracking task requires tracking the arbitrary target in consecutive video frames. Recently, several attempts have been made to develop the template-free models to attain generality. However, the current template-free paradigm only estimates the displacement to approximate the motion of the object. The displacement is insufficient to represent complex bounding box transformation, including scaling and rotation. We argue that the coarse-grained representation of object motion limits the performance of current template-free approaches. In this paper, we explore the finer-grained motion estimation to improve the accuracy of the template-free model. In respect of the image space, our method estimates the transformation for each pixel in the image. Concern on the motion representation, we represent the motion by the transformation parameterized by displacement, scaling, and rotation. By applying the differential vector operators on the optical flow, our approach estimates both displacement, scaling, and rotation for each pixel in a unified theory. To the best of our knowledge, we are the first work to model the displacement, scaling, and rotation in a unified theory with the optical flow. To further improve the localization accuracy, we develop the appearance branch to introduce the appearance information into our model. Furthermore, to suppress optical flow estimation failure samples during training, we propose a novel loss function Limited L1. The experiment shows our model FGTrack achieves state-of-the-art performance on both NFS and VOT2017 datasets.

****

Adaptive Aggregation of Arbitrary Online Trackers with a Regret Bound
Heon Song, Daiki Suehiro, Seiichi Uchida; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 681-689

We propose an online visual-object tracking method that is robust even in an adversarial environment, where various disturbances may occur on the target appearance, etc. The proposed method is based on a delayed-Hedge algorithm for aggregating multiple arbitrary online trackers with adaptive weights. The robustness in the tracking performance is guaranteed theoretically in term of "regret" by the property of the delayed-Hedge algorithm. Roughly speaking, the proposed method can achieve a similar tracking performance as the best one among all the trackers to be aggregated in an adversarial environment. The experimental study on various tracking tasks shows that the proposed method could achieve state-of-the-art performance by aggregating various online trackers.

****

Multiple Object Forecasting: Predicting Future Object Locations in Diverse Environments
Oliver Styles, Victor Sanchez, Tanaya Guha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 690-699

This paper introduces the problem of multiple object forecasting (MOF), in which the goal is to predict future bounding boxes of tracked objects. In contrast to existing works on object trajectory forecasting which primarily consider the problem from a birds-eye perspective, we formulate the problem from an object-level perspective and call for the prediction of full object bounding boxes, rather than trajectories alone. Towards solving this task, we introduce the Citywalks dataset, which consists of over 200k high-resolution video frames. Citywalks comprises of footage recorded in 21 cities from 10 European countries in a variety of weather conditions and over 3.5k unique pedestrian trajectories. For evaluatio n, we adapt existing trajectory forecasting methods for MOF and confirm cross-dataset generalizability on the MOT-17 dataset without fine-tuning. Finally, we present STED, a novel encoder-decoder architecture for MOF. STED combines visual and temporal features to model both object-motion and ego-motion, and outperforms existing approaches for MOF. Code & dataset link: https://github.com/olly-styles/Multiple-Object-Forecasting

****

Real-time Visual Object Tracking with Natural Language Description

Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, Stan Sclaroff; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 700-709

In this work, we argue that conditioning on the natural language (NL) description of a target provides information for longer-term invariance, and thus helps cope with typical tracking challenges. However, deriving a formulation to combine the strengths of appearance-based tracking with the language modality is not straightforward. Therefore, we propose a novel deep tracking-by-detection formulation that can take advantage of NL descriptions. Regions that are related to the given NL description are generated by a proposal network during the detection stage of the tracker. Our LSTM based tracker then predicts the update of the target from regions proposed by the NL based detection stage. Our method runs at over 30 fps on a single GPU. In benchmarks, our method is competitive with state of the art trackers that employ bounding boxes for initialization, while it outperforms all other trackers on targets given unambiguous and precise language annotations. When conditioned on NL descriptions only, our model doubles the performance of the previous best attempt.

****

Inverse Rectification for Efficient Procam Pattern Correspondence

Yubo Qiu, Jonathon Malcolm, Sheikh Ziauddin, Michael Greenspan, Abhay Vatoo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 710-718

A method called inverse rectification, is proposed which facilitates the establishment of correspondences across a projected pattern and an acquired image. A pattern of features comprising vertical dashes is warped by the inverse of the rectifying homography of the projector-camera pair, prior to projection. This warping imparts upon the system the property that projected features will fall on distinct conjugate epipolar lines of the rectified projector and acquired camera images. This reduces the correspondence search to a trivial constant-time table lookup once a feature is found in the camera image, and leads to robust, accurate, and extremely efficient disparity calculations. A projector-camera range sensor is developed based on this method, and is shown experimentally to be effective, with bandwidth exceeding some existing consumer-level range sensors.

****

Graph Networks for Multiple Object Tracking

Jiahe Li, Xu Gao, Tingting Jiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 719-728

Multiple object tracking (MOT) task requires reasoning the states of all targets and associating these targets in a global way. However, existing MOT methods mostly focus on the local relationship among objects and ignore the global relationship. Some methods formulate the MOT problem as a graph optimization problem. However, these methods are based on static graphs, which are seldom updated. To solve these problems, we design a new near-online MOT method with an end-to-end graph network. Specifically, we design an appearance graph network and a motion graph network to capture the appearance and the motion similarity separately. The updating mechanism is carefully designed in our graph network, which means that nodes, edges and the global variable in the graph can be updated. The global variable can capture the global relationship to help tracking. Finally, a strategy to handle missing detections is proposed to remedy the defect of the detectors. Our method is evaluated on both the MOT16 and the MOT17 benchmarks, and experimental results show the encouraging performance of our method.

****

Training with Noise Adversarial Network: A Generalization Method for Object Detection on Sonar Image

Qixiang Ma, Longyu Jiang, Wenxue Yu, Rui Jin, Zhixiang Wu, Fangjin Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 729-738

Object detection tasks for sonar image confront two major challenges, scarcity of dataset and perturbation of noise, which cause overfitting to models. The state-of-the-art object detection designed for optical images cannot address the issues because of the inherent differentiation between the optical image and sonar image. To tackle this problem, in this paper, we propose an adversarial training method to generalize the detector by introducing perturbation with specific noise property of sonar images during training stage. We design a sideway network which we name Noise Adversarial Network (NAN). The NAN is embedded into the state-of-the-art detector to generate adversarial examples which serve as assistant decision-making items to predict both class and bounding box, aiming to improve the generalization and noise robustness of the detector. To provide prior knowledge of noise perturbation to NAN, we also design a Noise Block (NB) for introducing noise in the upstream layers, which further improves noise robustness. Following the Faster R-CNN framework, the results of our experiments indicate a 8.9% mAP boost on our sonar image dataset. The detector equipped with NAN and NB also outperforms the baseline on noised test sets. Furthermore, it gains a 2.4% mAP boost on the optical image dataset PASCAL VOC 2007.

****

Active Adversarial Domain Adaptation

Jong-Chyi Su, Yi-Hsuan Tsai, Kihyuk Sohn, Buyu Liu, Subhransu Maji, Manmohan Chandraker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 739-748

We propose an active learning approach for transferring representations across domains. Our approach, active adversarial domain adaptation (AADA), explores a duality between two related problems: adversarial domain alignment and importance sampling for adapting models across domains. The former uses a domain discriminative model to align domains, while the latter utilizes the model to weigh samples to account for distribution shifts. Specifically, our importance weight promotes unlabeled samples with large uncertainty in classification and diversity compared to labeled examples, thus serving as a sample selection scheme for active learning. We show that these two views can be unified in one framework for domain adaptation and transfer learning when the source domain has many labeled examples while the target domain does not. AADA provides significant improvements over fine-tuning based approaches and other sampling methods when the two domains are closely related. Results on challenging domain adaptation tasks such as object detection demonstrate that the advantage over baseline approaches is retained even after hundreds of examples being actively annotated.

****

Progressive Domain Adaptation for Object Detection

Han-Kai Hsu, Chun-Han Yao, Yi-Hsuan Tsai, Wei-Chih Hung, Hung-Yu Tseng, Maneesh Singh, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 749-757

Recent deep learning methods for object detection rely on a large amount of bounding box annotations. Collecting these annotations is laborious and costly, yet supervised models do not generalize well when testing on images from a different distribution. Domain adaptation provides a solution by adapting existing labels to the target testing data. However, a large gap between domains could make adaptation a challenging task, which leads to unstable training processes and sub-optimal results. In this paper, we propose to bridge the domain gap with an intermediate domain and progressively solve easier adaptation subtasks. This intermediate domain is constructed by translating the source images to mimic the ones in the target domain. To tackle the domain-shift problem, we adopt adversarial learning to align distributions at the feature level. In addition, a weighted task loss is applied to deal with unbalanced image quality in the intermediate domain. Experimental results show that our method performs favorably against the state

-of-the-art method in terms of the performance on the target domain.

****

Boosting Standard Classification Architectures Through a Ranking Regularizer

Ahmed Taha, Yi-Ting Chen, Teruhisa Misu, Abhinav Shrivastava, Larry Davis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 758-766

We employ triplet loss as a feature embedding regularizer to boost classification performance. Standard architectures, like ResNet and Inception, are extended to support both losses with minimal hyper-parameter tuning. This promotes generality while fine-tuning pretrained networks. Triplet loss is a powerful surrogate for recently proposed embedding regularizers. Yet, it is avoided due to large batch-size requirement and high computational cost. Through our experiments, we re-assess these assumptions. During inference, our network supports both classification and embedding tasks without any computational overhead. Quantitative evaluation highlights a steady improvement on five fine-grained recognition datasets. Further evaluation on an imbalanced video dataset achieves significant improvement. Triplet loss brings feature embedding capabilities like nearest neighbor to classification models. Code available at http://bit.ly/2LNYEqL

****

Overlap Sampler for Region-Based Object Detection

Joya Chen, Bin Luo, Qi Wu, Jia Chen, Xuezheng Peng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 767-775

The top accuracy of object detection to date is led by region-based approaches, where the per-region stage is responsible for recognizing proposals generated by the region proposal network. In that stage, sampling heuristics (e.g., OHEM, IoU-balanced sampling) is always applied to select a part of examples during training. But nowadays, existing samplers ignore the overlaps among examples, which may result in some low-quality predictions preserved. To mitigate the issue, we propose Overlap Sampler that selects examples according to the overlaps among examples, which enables the training to focus on the important examples. Benefitted from it, the Faster R-CNN could obtain impressively 1.5 points higher Average Precision (AP) on the challenging COCO benchmark, a state-of-the-art result among existing samplers for region-based detectors. Moreover, the proposed sampler also yields considerable improvements for the instance segmentation task. Our code is released at https://github.com/ChenJoya/overlap-sampler.

****

A one-and-half stage pedestrian detector

Ujjwal Ujjwal, Aziz Dziri, Bertrand Leroy, Francois Bremond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 776-785

Pedestrian detection is a specific instance of the more general problem of object detection in computer vision. A balance between detection accuracy and speed is a desirable trait for pedestrian detection systems in many applications such as self-driving cars. In this paper, we follow the wisdom of " and less is often more" to achieve this balance. We propose a lightweight mechanism based on semantic segmentation to reduce the number of anchors to be processed. We furthermore unify this selection with the intra-anchor feature pooling strategy adopted in high performance two-stage detectors such as Faster-RCNN. Such astrategy is avoided in one-stage detectors like SSD in favourof faster inference but at the cost of reducing the accuracy vis-`a-vis two-stage detectors. However our anchor selection renders it practical to use feature pooling without giving up the inference speed. Our proposed approach succeeds in detecting pedestrians with state-of-art performance on caltech-reasonable and ciypersons datasets with inference speeds of 32fps.

****

Model-Agnostic Metric for Zero-Shot Learning

Jiayi Shen, Haochen Wang, Anran Zhang, Qiang Qiu, Xiantong Zhen, Xianbin Cao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 786-795

Zero-shot Learning (ZSL) aims to learn a classifier to recognize unseen categories without training samples. Most ZSL works based on embedding models handle the visual space and the semantic space through a common metric space and then apply a simple nearest neighbor search which directly leads to the hubness problem, one of the main challenges of ZSL. Contrary to recent works, whose conclusions about hubs are drawn based on Euclidean and specific models like ridge regression, we adopt cosine metric and for the first time prove cosine is model-agnostic to alleviate the hubness problem in ZSL. Assuming that the normalized mapped semantic vectors follow a uniform distribution, we provide theoretical analysis which demonstrates that hubs can be better reduced with a higher-dimensional cosine metric space. Moreover, we introduce a diversity-based regularizer with the cosine metric which underpins the assumption about the uniform distribution and further improves the model's discriminative ability. Extensive experiments on five benchmark datasets and large-scale Imagenet dataset show that our method can consistently improve the performance, surpassing previous embedding methods by large margins.

****

Intelligent Image Collection: Building the Optimal Dataset

Matthew Gwilliam, Ryan Farrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 796-805

Key recognition tasks such as fine-grained visual categorization (FGVC) have benefited from increasing attention among computer vision researchers. The development and evaluation of new approaches relies heavily on benchmark datasets; such datasets are generally built primarily with categories that have images readily available, omitting categories with insufficient data. This paper takes a step back and rethinks dataset construction, focusing on intelligent image collection driven by: (i) the inclusion of all desired categories, and, (ii) the recognition performance on those categories. Based on a small, author-provided initial dataset, the proposed system recommends which categories the authors should prioritize collecting additional images for, with the intent of optimizing overall categorization accuracy. We show that mock datasets built using this method outperform datasets built without such a guiding framework. Additional experiments give prospective dataset creators intuition into how, based on their circumstances and goals, a dataset should be constructed.

****

Internet of Things (IoT) Discovery Using Deep Neural Networks

Ephraim Lo, JoHannah Kohl; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 806-814

We present a novel approach to Internet of Things (IoT) discovery using Deep Neural Network (DNN) based object detection. Traditional methods of IoT discovery are based on either manual or automated monitoring of predetermined channel frequencies. Our method takes the spectrogram images that a human analyst visually scans for manual spectrum exploration and applies the state-of-the-art You Only Look Once (YOLO) object detection algorithm to detect and localize signal objects in time and frequency. We focus specifically on the class of signals that employ the Long Range (LoRa) modulation scheme, which uses chirp spread spectrum technology to provide high network efficiency and robustness against both in- and out-of-band interference. Our detection system is designed with scalability for real or near real-time processing capabilities and achieves 81.82% mAP in real-time on a fourth generation mobile Intel CPU without GPU support. Lastly, we present preliminary detection results for other IoT signals including Zigbee, Bluetooth, and Wi-Fi.

****

Propose-and-Attend Single Shot Detector

Ho-Deok Jang, Sanghyun Woo, Philipp Benz, Jinsun Park, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 815-824

We present a simple yet effective prediction module for a one-stage detector. The main process is conducted in a coarse-to-fine manner. First, the module roughly adjusts the default boxes to well capture the extent of target objects in an i mage. Second, given the adjusted boxes, the module aligns the receptive field of the convolution filters accordingly, not requiring any embedding layers. Both steps build a propose-and-attend mechanism, mimicking two-stage detectors in a highly efficient manner. To verify its effectiveness, we apply the proposed module to a basic one-stage detector SSD. We empirically show that our module significantly lifts the detection accuracy with marginal parameter overhead. Our final model achieves an accuracy comparable to that of state-of-the-art detectors while using a fraction of their model parameter and computational overheads. Moreover, we found that the proposed module has two strong applications. 1) The module can be successfully integrated into a lightweight backbone, further pushing the efficiency of the one-stage detector. 2) The module also allows train-from-scratch without relying on any sophisticated base networks as previous methods do.

****

Local Binary Pattern Networks

Jeng-Hau Lin, Justin Lazarow, Andrew Yang, Dezhi Hong, Rajesh Gupta, Zhuowen Tu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 825-834

Emerging edge devices such as sensor nodes are increasingly being tasked with non-trivial tasks related to sensor data processing and even application-level inferences from this sensor data. These devices are, however, extraordinarily resource-constrained in terms of CPU power (often Cortex M0-3 class CPUs), available memory (in few KB to MBytes), and energy. Under these constraints, we explore a novel approach to character recognition using local binary pattern networks, or LBPNet, that can learn and perform bit-wise operations in an end-to-end fashion. LBPNet has its advantage for characters whose features are composed of structured strokes and distinctive outlines. LBPNet uses local binary comparisons and random projections in place of conventional convolution (or approximation of convolution) operations, providing an important means to improve memory efficiency as well as inference speed. We evaluate LBPNet on a number of character recognition benchmark datasets as well as several object classification datasets and demonstrate its effectiveness and efficiency.

****

Leveraging Filter Correlations for Deep Model Compression

Pravendra Singh, Vinay Kumar Verma, Piyush Rai, Vinay Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 835-844

We present a filter correlation based model compression approach for deep convolutional neural networks. Our approach iteratively identifies pairs of filters with the largest pairwise correlations and drops one of the filters from each such pair. However, instead of discarding one of the filters from each such pair naively, the model is re-optimized to make the filters in these pairs maximally correlated, so that discarding one of the filters from the pair results in minimal information loss. Moreover, after discarding the filters in each round, we further finetune the model to recover from the potential small loss incurred by the compression. We evaluate our proposed approach using a comprehensive set of experiments and ablation studies. Our compression method yields state-of-the-art FLOPs compression rates on various benchmarks, such as LeNet-5, VGG-16, and ResNet-50,56, while still achieving excellent predictive performance for tasks such as object detection on benchmark datasets.

****

360-Indoor: Towards Learning Real-World Objects in 360deg Indoor Equirectangular Images

Shih-Han Chou, Cheng Sun, Wen-Yen Chang, Wan-Ting Hsu, Min Sun, Jianlong Fu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 845-853

While there are several widely used object detection datasets, current computer vision algorithms are still limited in conventional images. Such images narrow our vision in a restricted region. On the other hand, 360deg images provide a thorough sight. In this paper, our goal is to provide a standard dataset to facilitate the vision and machine learning communities in 360deg domain. To facilitate the research, we present a real-world 360deg panoramic object detection dataset, 360-Indoor, which is a new benchmark for visual object detection and class recognition in 360deg indoor images. It is achieved by gathering images of complex indoor scenes containing common objects and the intensive annotated bounding field-of-view. In addition, 360-Indoor has several distinct properties: (1) the largest category number (37 labels in total). (2) the most complete annotations on average (27 bounding boxes per image). The selected 37 objects are all common in indoor scene. With around 3k images and 90k labels in total, 360-Indoor achieves the largest dataset for detection in 360deg images. In the end, extensive experiments on the state-of-the-art methods for both classification and detection are provided. We will release this dataset in the near future.

****

Regularize, Expand and Compress: NonExpansive Continual Learning

Jie Zhang, Junting Zhang, Shalini Ghosh, Dawei Li, Jingwen Zhu, Heming Zhang, Yalin Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 854-862

Continual learning (CL), the problem of lifelong learning where tasks arrive in sequence, has attracted increasing attention in the computer vision community lately. The goal of CL is to learn new tasks while maintaining the performance on the previously learned tasks. There are two major obstacles for CL of deep neural networks: catastrophic forgetting and limited model capacity. Inspired by the recent breakthroughs in automatically learning good neural network architectures, we develop a nonexpansive AutoML framework for CL termed Regularize, Expand and Compress (REC) to solve the above issues. REC is a unified framework with three highlights: 1) a novel regularized weight consolidation (RWC) algorithm to avoid forgetting, where accessing the data seen in the previously learned tasks is not required; 2) an automatic neural architecture search (AutoML) engine to expand the network to increase model capability; 3) smart compression of the expanded model after a new task is learned to improve the model efficiency. The experimental results on four different image recognition datasets demonstrate the superior performance of the proposed REC over other CL algorithms.

****

Synthetic Examples Improve Generalization for Rare Classes

Sara Beery, Yang Liu, Dan Morris, Jim Piavis, Ashish Kapoor, Neel Joshi, Markus Meister, Pietro Perona; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 863-873

The ability to detect and classify rare occurrences in images has important applications -- for example, counting rare and endangered species when studying biodiversity, or detecting infrequent traffic scenarios that pose a danger to self-driving cars. Few-shot learning is an open problem: current computer vision systems struggle to categorize objects they have seen only rarely during training, and collecting a sufficient number of training examples of rare events is often challenging and expensive, and sometimes outright impossible. We explore in depth an approach to this problem: complementing the few available training images with ad-hoc simulated data.   Our testbed is animal species classification, which has a real-world long-tailed distribution. We analyze the effect of different axes of variation in simulation, such as pose, lighting, model, and simulation method, and we prescribe best practices for efficiently incorporating simulated data for real-world performance gain. Our experiments reveal that synthetic data can considerably reduce error rates for classes that are rare, that as the amount of simulated data is increased, accuracy on the target class improves, and that high variation of simulated data provides maximum performance gain.

****

CANZSL: Cycle-Consistent Adversarial Networks for Zero-Shot Learning from Natural Language

Zhi Chen, Jingjing Li, Yadan Luo, Zi Huang, Yang Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 874-883

Existing methods using generative adversarial approaches for Zero-Shot Learning (ZSL) aim to generate realistic visual features from class semantics by a single

generative alignment, which is highly under-constrained. As a result, the previous methods cannot guarantee that the generated visual features can truthfully reflect the corresponding semantics. To address this issue, we propose a novel method named Cycle-consistent Adversarial Networks for Zero-Shot Learning (CANZSL). It encourages a visual feature generator to synthesize realistic visual features from semantics, and then inversely translate back the synthesized visual features to the corresponding semantic space by a semantic feature generator. Furthermore, in this paper a more challenging and practical ZSL problem is considered where the original semantics are from natural language with irrelevant words instead of clean semantics, which are widely used in previous work. Specifically, a multi-modal consistent bidirectional generative adversarial model is trained to handle unseen instances by suppressing noise in the natural language. A forward one-to-many mapping from the class level descriptions to the visual features is coupled with an inverse many-to-one mapping from the visual space to the semantic space. Thus, a multi-modal cycle-consistency loss between the synthesized semantic representations and the ground truth can be learned and leveraged to enforce the generated semantic features to approximate to the real distribution in semantic space. Extensive experiments are conducted to demonstrate that our method consistently outperforms state-of-the-art approaches on natural language-based zero-shot learning tasks.
****
Accuracy Booster: Performance Boosting using Feature Map Re-calibration
Pravendra Singh, PRATIK MAZUMDER, Vinay Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 884-893
Convolution Neural Networks (CNN) have been extremely successful in solving intensive computer vision tasks. The convolutional filters used in CNNs have played a major role in this success, by extracting useful features from the inputs. Recently researchers have tried to boost the performance of CNNs by re-calibrating the feature maps produced by these filters, e.g., Squeeze-and-Excitation Networks (SENets). These approaches have achieved better performance by Exciting up the important channels or feature maps while diminishing the rest. However, in the process, architectural complexity has increased. We propose an architectural block that introduces much lower complexity than the existing methods of CNN performance boosting while performing significantly better than them. We carry out experiments on the CIFAR, ImageNet and MS-COCO datasets, and show that the proposed block can challenge the state-of-the-art results. Our method boosts the ResNet-50 architecture to perform comparably to the ResNet-152 architecture, which is a three times deeper network, on classification. We also show experimentally that our method is not limited to classification but also generalizes well to other tasks such as object detection.
****
Generating Positive Bounding Boxes for Balanced Training of Object Detectors
Kemal Oksuz, Baris Can Cam, Emre Akbas, Sinan Kalkan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 894-903
Two-stage deep object detectors generate a set of regions-of-interest (RoIs) in the first stage, then, in the second stage, identify objects among the proposed RoIs that sufficiently overlap with a ground truth (GT) box. The second stage is known to suffer from a bias towards RoIs that have low intersection-over-union (IoU) with the associated GT boxes. To address this issue, we first propose a sampling method to generate bounding boxes (BB) that overlap with a given reference box more than a given IoU threshold. Then, we use this BB generation method to develop a positive RoI (pRoI) generator that, for the second stage, produces RoIs following any desired spatial or IoU distribution. We show that our pRoI generator is able to simulate other sampling methods for positive examples such as hard example mining and prime sampling. Using our generator as an analysis tool, we show that (i) IoU imbalance has an adverse effect on performance, (ii) hard positive example mining improves the performance only for certain input IoU distributions, and (iii) the imbalance among the foreground classes has an adverse effect on performance and that it can be alleviated at the batch level. Finally, w

e train Faster R-CNN using our pRoI generator and, compared to conventional training, obtain better or on-par performance for low IoUs and significant improvements when trained for higher IoUs for Pascal VOC and MS COCO datasets. The code is available at: https://github.com/kemaloksuz/BoundingBoxGenerator

****

Towards Learning Affine-Invariant Representations via Data-Efficient CNNs
Wenju Xu, Guanghui Wang, Alan Sullivan, Ziming Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 904-913

In this paper we propose integrating a priori knowledge into both design and training of convolutional neural networks (CNNs) to learn object representations that are invariant to affine transformations (i.e. translation, scale, rotation). Accordingly we propose a novel multi-scale maxout CNN and train it end-to-end with a novel rotation-invariant regularizer. This regularizer aims to enforce the weights in each 2D spatial filter to approximate circular patterns. In this way, we manage to handle affine transformations in training using convolution, multi-scale maxout, and circular filters. Empirically we demonstrate that such knowledge can significantly improve the data-efficiency as well as generalization and robustness of learned models. For instance, on the Traffic Sign data set and trained with only 10 images per class, our method can achieve 84.15% that outperforms the state-of-the-art by 29.80% in terms of test accuracy.

****

Is Pruning Compression?: Investigating Pruning Via Network Layer Similarity
Cody Blakeney, Yan Yan, Ziliang Zong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 914-922

Unstructured neural network pruning is an effective technique that can significantly reduce theoretical model size, computation demand and energy consumption of large neural networks without compromising accuracy. However, a number of fundamental questions about pruning are not answered yet. For example, do the pruned neural networks contain the same representations as the original network? Is pruning a compression or evolution process? Does pruning only work on trained neural networks? What is the role and value of the uncovered sparsity structure? In this paper, we strive to answer these questions by analyzing three unstructured pruning methods (magnitude based pruning, post-pruning re-initialization, and random sparse initialization). We conduct extensive experiments using the Singular Vector Canonical Correlation Analysis (SVCCA) tool to study and contrast layer representations of pruned and original ResNet, VGG, and ConvNet models. We have several interesting observations: 1) Pruned neural network models evolve to substantially different representations while still maintaining similar accuracy. 2) Initialized sparse models can achieve reasonably good accuracy compared to well-engineered pruning methods. 3) Sparsity structures discovered by pruning models are not inherently important or useful.

****

Transductive Zero-Shot Learning for 3D Point Cloud Classification
Ali Cheraghian, Shafin Rahman, Dylan Campbell, Lars Petersson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 923-933

Zero-shot learning, the task of learning to recognize new classes not seen during training, has received considerable attention in the case of 2D image classification. However despite the increasing ubiquity of 3D sensors, the corresponding 3D point cloud classification problem has not been meaningfully explored and introduces new challenges. This paper extends, for the first time, transductive Zero-Shot Learning (ZSL) and Generalized Zero-Shot Learning (GZSL) approaches to the domain of 3D point cloud classification. To this end, a novel triplet loss is developed that takes advantage of unlabeled test data. While designed for the task of 3D point cloud classification, the method is also shown to be applicable to the more common use case of 2D image classification. An extensive set of experiments is carried out, establishing state-of-the-art for ZSL and GZSL in the 3D point cloud domain, as well as demonstrating the applicability of the approach to the image domain.

****

Wide Hidden Expansion Layer for Deep Convolutional Neural Networks

Min Wang, Baoyuan Liu, Hassan Foroosh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 934-942

Non-linearity is an essential factor contributing to the success of deep convolutional neural networks. Increasing the non-linearity in the network will enhance the network's learning capability, attributing to better performance. We present a novel Wide Hidden Expansion (WHE) layer that can significantly increase (by an order of magnitude) the number of activation functions in the network, with very little increase of computational complexity and memory consumption. It can be flexibly embedded with different network architectures to boost the performance of the original networks. The WHE layer is composed of a wide hidden layer, in which each channel only connects with two input channels and one output channel. Before connecting to the output channel, each intermediate channel in the WHE layer is followed by one activation function. In this manner, the number of activation functions can grow along with the number of channels in the hidden layer. We apply the WHE layer to ResNet, WideResNet, SENet, and MobileNet architectures and evaluate on ImageNet, CIFAR-100, and Tiny ImageNet dataset. On the ImageNet dataset, models with the WHE layer can achieve up to 2.01% higher Top-1 accuracy than baseline models, with less than 4% computation increase and less than 2% more parameters. On CIFAR-100 and Tiny ImageNet, when applying the WHE layer to ResNet models, it demonstrates consistent improvement in the accuracy of the networks. Applying the WHE layer to ResNet backbone of the CenterNet object detection model can also boost its performance on COCO and Pascal VOC datasets.

****

Learning from THEODORE: A Synthetic Omnidirectional Top-View Indoor Dataset for Deep Transfer Learning

Tobias Scheck, Roman Seidel, Gangolf Hirtz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 943-952

Recent work about synthetic indoor datasets from perspective views has shown significant improvements of object detection results with Convolutional Neural Networks(CNNs). In this paper, we introduce THEODORE: a novel, large-scale indoor dataset containing 100,000 high- resolution diversified fisheye images with 14 classes. To this end, we create 3D virtual environments of living rooms, different human characters and interior textures. Beside capturing fisheye images from virtual environments we create annotations for semantic segmentation, instance masks and bounding boxes for object detection tasks. We compare our synthetic dataset to state of the art real-world datasets for omnidirectional images. Based on MS COCO weights, we show that our dataset is well suited for fine-tuning CNNs for object detection. Through a high generalization of our models by means of image synthesis and domain randomization we reach an AP up to 0.84 for class person on High-Definition Analytics dataset.

****

TKD: Temporal Knowledge Distillation for Active Perception

Mohammad Farhadi Bajestani, Yezhou Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 953-962

Deep neural network-based methods have been proved to achieve outstanding performance on object detection and classification tasks. Despite the significant performance improvement using the deep structures, they still require prohibitive runtime to process images and maintain the highest possible performance for real-time applications. Observing the phenomenon that human visual system (HVS) relies heavily on the temporal dependencies among frames from the visual input to conduct recognition efficiently, we propose a novel framework dubbed as TKD: temporal knowledge distillation. This framework distills the temporal knowledge from a heavy neural network-based model over selected video frames (the perception of the moments) to a light-weight model. To enable the distillation, we put forward two novel procedures: 1) a Long-short Term Memory (LSTM)-based key frame selection method; and 2) a novel teacher-bounded loss design. To validate our approach, we conduct comprehensive empirical evaluations using different object detection methods over multiple datasets including Youtube Objects and Hollywood scene da taset. Our results show consistent improvement in accuracy-speed trade-offs for object detection over the frames of the dynamic scene, compared to other modern object recognition methods. It can maintain the desired accuracy with the throughput of around 220 images per second. Implementation: https://github.com/mfarhadi/TKD-Cloud.

****

SketchTransfer: A New Dataset for Exploring Detail-Invariance and the Abstractions Learned by Deep Networks

Alex Lamb, Sherjil Ozair, Vikas Verma, David Ha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 963-972

Deep networks have achieved excellent results in perceptual tasks, yet their ability to generalize to variations not seen during training has come under increasing scrutiny. In this work we focus on their ability to have invariance towards the presence or absence of details. For example, humans are able to watch cartoons, which are missing many visual details, without being explicitly trained to do so. As another example, 3D rendering software is a relatively recent development, yet people are able to understand such rendered scenes even though they are missing details (consider a film like Toy Story). This capability goes beyond visual data: humans are easily able to recognize isolated melodies from musical pieces when heard for the first time, even if the only piece they've listened to previously is from an orchestra. Thus the failure of machine learning algorithms to do this indicates a significant gap in generalization between human abilities and the abilities of deep networks. We propose a dataset that will make it easier to study the detail-invariance problem concretely. We produce a concrete task for this: SketchTransfer, and we show that state-of-the-art domain transfer algorithms still struggle with this task. The state-of-the-art technique which achieves over 95% on MNIST \xrightarrow SVHN transfer only achieves 59% accuracy on the SketchTransfer task, which is much better than random (11% accuracy) but falls short of the 87% accuracy of a classifier trained directly on labeled sketches. This indicates that this task is approachable with today's best methods but has substantial room for improvement.

****

SVIRO: Synthetic Vehicle Interior Rear Seat Occupancy Dataset and Benchmark

Steve Dias Da Cruz, Oliver Wasenmuller, Hans-Peter Beise, Thomas Stifter, Didier Stricker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 973-982

We release SVIRO, a synthetic dataset for sceneries in the passenger compartment of ten different vehicles, in order to analyze machine learning-based approaches for their generalization capacities and reliability when trained on a limited number of variations (e.g. identical backgrounds and textures, few instances per class). This is in contrast to the intrinsically high variability of common benchmark datasets, which focus on improving the state-of-the-art of general tasks. Our dataset contains bounding boxes for object detection, instance segmentation masks, keypoints for pose estimation and depth images for each synthetic scenery as well as images for each individual seat for classification. The advantage of our use-case is twofold: The proximity to a realistic application to benchmark new approaches under novel circumstances while reducing the complexity to a more tractable environment, such that applications and theoretical questions can be tested on a more challenging dataset as toy problems. The data and evaluation server are available under https://sviro.kl.dfki.de.

****

Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization

saurabh desai, Harish Guruprasad Ramaswamy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 983-991

In response to recent criticism of gradient-based visualization techniques, we propose a new methodology to generate visual explanations for deep Convolutional Neural Networks (CNN) - based models. Our approach - Ablation-based Class Activation Mapping (Ablation CAM) uses ablation analysis to determine the importance (weights) of individual feature map units w.r.t. class. Further, this is used to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Our objective and subjective evaluations show that this gradient-free approach works better than state-of-the-art Grad-CAM technique. Moreover, further experiments are carried out to show that Ablation-CAM is class discriminative as well as can be used to evaluate trust in a model.

****

Supervised and Unsupervised Learning of Parameterized Color Enhancement
Yoav Chai, Raja Giryes, Lior Wolf; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 992-1000

We treat the problem of color enhancement as an image translation task, which we tackle using both supervised and unsupervised learning. Unlike traditional image to image generators, our translation is performed using a global parameterized color transformation instead of learning to directly map image information. In the supervised case, every training image is paired with a desired target image and a convolutional neural network (CNN) learns from the expert retouched images the parameters of the transformation. In the unpaired case, we employ two-way generative adversarial networks (GANs) to learn these parameters and apply a circularity constraint. We achieve state-of-the-art results compared to both supervised (paired data) and unsupervised (unpaired data) image enhancement methods on the MIT-Adobe FiveK benchmark. Moreover, we show the generalization capability of our method, by applying it on photos from the early 20th century and to dark video frames.

****

ADNet: Adaptively Dense Convolutional Neural Networks
Mingjie Wang, Hao Cai, Xin Huang, Minglun Gong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1001-1010

Convolutional neural networks (CNNs) have demonstrated great success in vision tasks. However, most existing architectures still suffer from low feature reuse efficiency. In this paper, we present a layer attention based Adaptively Dense Network (ADNet) by adaptively determining the reuse status of hierarchical preceding features. Specifically, a dense residual aggregation strategy is developed to fuse multi-level internal representations in an effective manner. Furthermore, a novel layer attention mechanism is proposed to explicitly model the interrelationship among layers to automatically adjust the density of the network. It is worth noting that existing ResNets and DenseNets are both special cases of our ADNet. Extensive experiments demonstrate that the proposed architecture consistently and indubitably achieves competitive results in accuracy on benchmark datasets (CIFAR10, CIFAR100, and SVHN), while at the same time remarkably reduces computational costs and memory space. Visualization and analysis on layer-wise attention further provide better understanding on the density of feature reuse in Deep Networks.

****

Exploring 3 R's of Long-term Tracking: Redetection, Recovery and Reliability
Shyamgopal Karthik, Abhinav Moudgil, Vineet Gandhi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1011-1020

Recent works have proposed several long term tracking benchmarks and highlight the importance of moving towards long-duration tracking to bridge the gap with application requirements. The current evaluation methodologies, however, do not focus on several aspects that are crucial in a long term perspective like Re-detection, Recovery, and Reliability. In this paper, we propose novel evaluation strategies for a more in-depth analysis of trackers from a long-term perspective. More specifically, (a) we test re-detection capability of the trackers in the wild by simulating virtual cuts, (b) we investigate the role of chance in the recovery of tracker after failure and (c) we propose a novel metric allowing visual inference on the ability of a tracker to track contiguously (without any failure) at a given accuracy. We present several original insights derived from an extensive set of quantitative and qualitative experiments.

****

The Overlooked Elephant of Object Detection: Open Set

Akshay Dhamija, Manuel Gunther, Jonathan Ventura, Terrance Boult; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1021-1030

Even though object detection is a popular area of research that has found considerable applications in the real world, it has some fundamental aspects that have never been formally discussed and experimented. One of the core aspects of evaluating object detectors has been the ability to avoid false detections. While major datasets like PASCAL VOC or MSCOCO extensively test the detectors on their ability to avoid false positives, they do not differentiate between their closed-set and open-set performance. Despite systems being trained to reject everything other than the classes of interest, unknown objects from the open world end up being incorrectly detected as known objects, often with very high confidence. This paper is the first to formalize the problem of open-set object detection and propose the first open-set object detection protocol. Moreover, the paper provides a new evaluation metric to analyze the performance of some state-of-the-art detectors and discusses their performance differences.

****

Probabilistic Object Detection: Definition and Evaluation

David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, Niko Suenderhauf; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1031-1040

We introduce Probabilistic Object Detection, the task of detecting objects in images and accurately quantifying the spatial and semantic uncertainties of the detections. Given the lack of methods capable of assessing such probabilistic object detections, we present the new Probability-based Detection Quality measure (PDQ). Unlike AP-based measures, PDQ has no arbitrary thresholds and rewards spatial and label quality, and foreground/background separation quality while explicitly penalising false positive and false negative detections. We contrast PDQ with existing mAP and moLRP measures by evaluating state-of-the-art detectors and a Bayesian object detector based on Monte Carlo Dropout. Our experiments indicate that conventional object detectors tend to be spatially overconfident and thus perform poorly on the task of probabilistic object detection. Our paper aims to encourage the development of new object detection approaches that provide detections with accurately estimated spatial and label uncertainties and are of critical importance for deployment on robots and embodied AI systems in the real world.

****

DeepPTZ: Deep Self-Calibration for PTZ Cameras

Chaoning Zhang, Francois Rameau, Junsik Kim, Dawit Mureja Argaw, Jean-Charles Bazin, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1041-1049

Rotating and zooming cameras, also called PTZ (Pan-Tilt-Zoom) cameras, are widely used in modern surveillance systems. While their zooming ability allows acquiring detailed images of the scene, it also makes their calibration more challenging since any zooming action results in a modification of their intrinsic parameters. Therefore, such camera calibration has to be computed online; this process is called self-calibration. In this paper, given an image pair captured by a PTZ camera, we propose a deep learning based approach to automatically estimate the focal length and distortion parameters of both images as well as the rotation angles between them. The proposed approach relies on a dual-Siamese structure, imposing bidirectional constraints. The proposed network is trained on a large-scale dataset automatically generated from a set of panoramas. Empirically, we demonstrate that our proposed approach achieves competitive performance with respect to both deep learning-based and traditional state-of-the art methods. Our code and model will be publicly available at https://github.com/ChaoningZhang/DeepPTZ.

****

Self-Orthogonality Module: A Network Architecture Plug-in for Learning Orthogonal Filters

Ziming Zhang, WENCHI MA, Yuanwei Wu, Guanghui Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1050-1059

In this paper, we investigate the empirical impact of or- thogonality regularization (OR) in deep learning, either solo or collaboratively. Recent works on OR showed some promis- ing results on the accuracy. In our ablation study, however, we do not observe such significant improvement from exist- ing OR techniques compared with the conventional training based on weight decay, dropout, and batch normalization. To identify the real gain from OR, inspired by the locality sensitive hashing (LSH) in angle estimation, we propose to introduce an implicit self-regularization into OR to push the mean and variance of filter angles in a network towards 90 * and 0 * simultaneously to achieve (near) orthogonality among the filters, without using any other explicit regular- ization. Our regularization can be implemented as an archi- tectural plug-in and integrated with an arbitrary network. We reveal that OR helps stabilize the training process and leads to faster convergence and better generalization.

****

Evaluation of Image Inpainting for Classification and Retrieval

Samuel Black, Somayeh Keshavarz, Richard Souvenir; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1060-1069

A common approach to censoring digital image content is masking the region(s) of interest with a solid color or pattern. In the case where the masked image will be used as input for classification or matching, the mask itself may impact the results. Recent work in image inpainting provides an alternative to masking by replacing the foreground with predicted background. In this paper, we perform an extensive evaluation of inpainting approaches to understand how well inpainted images can serve as proxies for the original in classification and retrieval. Results indicate that the metrics typically used to evaluate inpainting performance (e.g., reconstruction accuracy) do not necessarily correspond to improved classification or retrieval, especially in the case of person-shaped masked regions.

****

Adversarial Examples for Edge Detection: They Exist, and they Transfer

Christian Cosgrove, Alan Yuille; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1070-1079

Convolutional neural networks have recently advanced the state of the art in many tasks including edge and object boundary detection. However, in this paper, we demonstrate that these edge detectors inherit a troubling property of neural networks: they can be fooled by adversarial examples. We show that adding small perturbations to an image causes HED, a CNN-based edge detection model, to fail to locate edges, to detect nonexistent edges, and even to hallucinate arbitrary configurations of edges. More importantly, we find that these adversarial examples blindly transfer to other CNN-based vision models. In particular, attacks on edge detection result in significant drops in accuracy in models trained to perform unrelated, high-level tasks like image classification and semantic segmentation.

****

Spatio-Temporal Pyramid Graph Convolutions for Human Action Recognition and Postural Assessment

Behnoosh Parsa, Athmanarayanan Lakshmi narayanan, Behzad Dariush; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1080-1090

Recognition of human actions and associated interactions with objects and the environment is an important problem in computer vision due to its potential applications in a variety of domains. Recently, graph convolutional networks that extract features from the skeleton have demonstrated promising performance. In this paper, we propose a novel Spatio-Temporal Pyramid Graph Convolutional Network (ST-PGN) for online action recognition for ergonomics risk assessment that enables the use of features from all levels of the skeleton feature hierarchy. The prop osed algorithm outperforms state-of-art action recognition algorithms tested on two public benchmark datasets typically used for postural assessment (TUM and UW-IOM). We also introduce a pipeline to enhance postural assessment methods with online action recognition techniques. Finally, the proposed algorithm is integrated with a traditional ergonomics risk index (REBA) to demonstrate the potential value for assessment of musculoskeletal disorders in occupational safety.

****

Towards Good Practice for CNN-Based Monocular Depth Estimation

Zhicheng Fang, Xiaoran Chen, Yuhua Chen, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1091-1100

Monocular depth estimation has gained increasing attention in recent years, and various techniques have been proposed to tackle this problem. In this work, we aim to provide a comprehensive study on the techniques widely used in monocular depth estimation, and examine their individual influence on the performance. More specifically, we provide a study on: 1) network architectures, including different combinations of encoders/decoders. 2) supervision losses, including fully supervised losses and self-supervised losses and 3) other practices such as input resolution. The experiments are conducted on two commonly used public datasets, KITTI and NYU Depth v2. We also provide an analysis on the errors produced by different models, to reveal the limitations of current methods. Furthermore, by a careful redesign, we present a model for depth estimation, which achieves competitive performance on KITTI and state-of-the-art performance on NYU Depth v2. Our code is publicly available at https://github.com/zenithfang/supervised_dispnet.

****

I-MOVE: Independent Moving Objects for Velocity Estimation

Jonathan Schwan, Akshay Dhamija, Terrance Boult; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1101-1110

We introduce I-MOVE, the first publicly available RGB-D/stereo dataset for estimating velocities of independently moving objects. Velocity estimation given RGB-D data is an unsolved problem. The I-MOVE dataset provides an opportunity for generalizable velocity estimation models to be created and have their performance be accurately and fairly measured. The dataset features various outdoor and indoor scenes of single and multiple moving objects. Compared to other datasets, I-MOVE is unique because the RGB-D data and speed for each object are supplied for a variety of different settings/environments, objects, and motions. The dataset includes training and test sequences captured from four different depth camera views and three 4K-stereo setups. The data are also time-synchronized with three Doppler radars to provide the magnitude of velocity ground truth. The I-MOVE dataset includes complex scenes from moving pedestrians via walking and biking to multiple rolling objects, all captured with the seven cameras, providing over 500 Depth/Stereo videos.

****

Real-time vehicle distance estimation using single view geometry

Ahmed Ali, Ali Hassan, Afsheen Rafaqat Ali, Hussam Ullah Khan, Wajahat Kazmi, Aamer Zaheer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1111-1120

Distance estimation is required for advanced driver as- sistance systems (ADAS) as well as self-driving cars. It is crucial for obstacle avoidance, tailgating detection and accident prevention. Currently, radars and lidars are pri- marily used for this purpose which are either expensive or offer poor resolution. Deep learning based depth or dis- tance estimation techniques require huge amount of data and ensuring domain invariance is a challenge. Therefore, in this paper, we propose a single view geometric approach which is lightweight and uses geometric features of the road lane markings for distance estimation that integrates well with the lane and vehicle detection modules of an existing ADAS. Our system introduces novelty on two fronts: (1) it uses cross-ratios of lane boundaries to estimate horizon (2) it determines an Inverse Perspective Mapping (IPM) and camera height from a known lane width and the detected horizon. Distances of the vehicles on the road are then cal- culated by back projecting image point to a ray in tersecting the reconstructed road plane. For evaluation, we used li- dar data as ground truth and compare the performance of our algorithm with radar as well as the state-of-the-art deep learning based monocular depth prediction algorithms. The results on three public datasets (Kitti, nuScenes and Lyft level 5) showed that the proposed system maintains a con- sistent RMSE between 6.10 to 7.31. It outperforms other algorithms on two of the datasets while on KITTI it falls behind one (supervised) deep learning method. Furthermore, it is computationally inexpensive and is data-domain invari- ant.

****

SmartOverlays: A Visual Saliency Driven Label Placement for Intelligent Human-Computer Interfaces

Srinidhi Hegde, Jitender Maurya, Aniruddha Kalkar, Ramya Hebbalaguppe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1121-1130

In augmented reality (AR), the computer generated labels assist in understanding a scene by addition of contextual information. However, naive label placement often results in clutter and occlusion impairing the effectiveness of AR visualization. For label placement, the main objectives to be satisfied are, non-occlusion to the scene of interest, the proximity of labels to the object, and, temporally coherent labels in a video/live feed. We present a novel method for the placement of labels corresponding to objects of interest in a video/live feed that satisfies the aforementioned objectives. Our proposed framework, SmartOverlays, first identifies the objects and generates corresponding labels using a YOLOv2 in a video frame; at the same time, Saliency Attention Model (SAM) learns eye fixation points that aid in predicting saliency maps; finally, computes Voronoi partitions of the video frame, choosing the centroids of objects as seed points, to place labels for satisfying the proximity constraints with the object of interest. In addition, our approach incorporates tracking the detected objects in a frame to facilitate temporal coherence between frames that enhances the readability of labels. We measure the effectiveness of SmartOverlays framework using three objective metrics: (a) Label Occlusion over Saliency (LOS), (b) temporal jitter metric to quantify jitter in the label placement, (c) computation time for label placement.

****

Class-incremental Learning via Deep Model Consolidation

Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, C.-C. Jay Kuo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1131-1140

Deep neural networks (DNNs) often suffer from "catastrophic forgetting" during incremental learning (IL) --- an abrupt degradation of performance on the original set of classes when the training objective is adapted to a newly added set of classes. Existing IL approaches tend to produce a model that is biased towards either the old classes or new classes, unless with the help of exemplars of the old data. To address this issue, we propose a class-incremental learning paradigm called Deep Model Consolidation (DMC), which works well even when the original training data is not available. The idea is to first train a separate model only for the new classes, and then combine the two individual models trained on data of two distinct set of classes (old classes and new classes) via a novel double distillation training objective. The two existing models are consolidated by exploiting publicly available unlabeled auxiliary data. This overcomes the potential difficulties due to unavailability of original training data. Compared to the state-of-the-art techniques, DMC demonstrates significantly better performance in image classification (CIFAR-100 and CUB-200) and object detection (PASCAL VOC 2007) in the single-headed IL setting.

****

Cooperative Initialization based Deep Neural Network Training

Pravendra Singh, Munender Varshney, Vinay Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1141-1150

Researchers have proposed various activation functions. These activation functio ns help the deep network to learn non-linear behavior with a significant effect on training dynamics and task performance. The performance of these activations also depends on the initial state of the weight parameters, i.e., different initial state leads to a difference in the performance of a network. In this paper, we have proposed a cooperative initialization for training the deep network using ReLU activation function to improve the network performance. Our approach uses multiple activation functions in the initial few epochs for the update of all sets of weight parameters while training the network. These activation functions cooperate to overcome their drawbacks in the update of weight parameters, which in effect learn better "feature representation" and boost the network performance later. Cooperative initialization based training also helps in reducing the overfitting problem and does not increase the number of parameters, inference (test) time in the final model while improving the performance. Experiments show that our approach outperforms various baselines and, at the same time, performs well over various tasks such as classification and detection. The Top-1 classification accuracy of the model trained using our approach improves by 2.8% for VGG-16 and 2.1% for ResNet-56 on CIFAR-100 dataset.

****

Self-Growing Spatial Graph Networks for Pedestrian Trajectory Prediction

Sirin Haddad, Siew-Kei Lam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1151-1159

Intelligent vehicles and social robots need to navigate in crowded environments while avoiding collisions with pedestrians. To achieve this, pedestrian trajectory prediction is essential. However, predicting pedestrians' trajectory in crowded environments is nontrivial as human-to-human interactions among the crowd participants influence their motion. In this work, we propose a novel end-to-end graph-centric gated learning model to estimate the existence of interactions between individuals. Accordingly, the model predicts pedestrians' future locations and velocities. Recent methods based on LSTM networks used thresholding techniques to define neighborhood boundaries and relationships. Other graph-structured methods grow edges in polynomial size. In contrast, our graph-based GRU network model employs an online data-driven criterion that can learn from interactions and grow connections between pedestrian nodes. The proposed model yields outperforming prediction accuracy over state-of-the-art works in two public datasets, i.e. Crowds and SDD.

****

ImaGINator: Conditional Spatio-Temporal GAN for Video Generation

Yaohui WANG, Piotr Bilinski, Francois Bremond, Antitza Dantcheva; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1160-1169

Generating human videos based on single images entails the challenging simultaneous generation of realistic and visual appealing appearance and motion. In this context, we propose a novel conditional GAN architecture, namely ImaGINator, which given a single image, a condition (label of a facial expression or action) and noise, decomposes appearance and motion in both latent and high level feature spaces, generating realistic videos. This is achieved by (i)a novel spatio-temporal fusion scheme, which generates dynamic motion, while retaining appearance throughout the full video sequence by transmitting appearance (originating from the single image) through all layers of the network. In addition, we propose (ii) a novel transposed (1+2)D convolution, factorizing the transposed 3D convolutional filters into separate transposed temporal and spatial components, which yields significantly gains in video quality and speed. We extensively evaluate our approach on the facial expression datasets MUG and UvA-NEMO, as well as on the action datasets NATOPS and Weizmann. We show that our approach achieves significantly better quantitative and qualitative results than the state-of-the-art. The source code and models are available under https://github.com/wyhsirius/ImaGINator.

****

One-to-one Mapping for Unpaired Image-to-image Translation

Zengming Shen, S. Kevin Zhou, Yifan Chen, Bogdan Georgescu, Xuqi Liu, Thoma s Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1170-1179

Recently image-to-image translation has attracted significant interests in the literature, starting from the successful use of the generative adversarial network (GAN), to the introduction of cyclic constraint, to extensions to multiple domains. However, in existing approaches, there is no guarantee that the mapping between two image domains is unique or one-to-one. Here we propose a self-inverse network learning approach for unpaired image-to-image translation. Building on top of CycleGAN, we learn a self-inverse function by simply augmenting the training samples by switching inputs and outputs during training. The outcome of such learning is a proven one-to-one mapping function. Our extensive experiments on a variety of detests, including cross-modal medical image synthesis, object transfiguration, and semantic labeling, consistently demonstrate clear improvement over the CycleGAN method both qualitatively and quantitatively. Especially our proposed method reaches the state-of-the-art result on the label to photo direction of the cityscapes benchmark dataset.

****

ViP: Virtual Pooling for Accelerating CNN-based Image Classification and Object Detection

Zhuo Chen, Jiyuan Zhang, Ruizhou Ding, Diana Marculescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1180-1189

In recent years, Convolutional Neural Networks (CNNs) have shown superior capability in visual learning tasks. While accuracy-wise CNNs provide unprecedented performance, they are also known to be computationally intensive and energy demanding for modern computer systems. In this paper, we propose Virtual Pooling (ViP), a model-level approach to improve speed and energy consumption of CNN-based image classification and object detection tasks, with a provable error bound. We show the efficacy of ViP through experiments on four CNN models, three representative datasets, both desktop and mobile platforms, and two visual learning tasks, i.e., image classification and object detection. For example, ViP delivers 2.1x speedup with less than 1.5% accuracy degradation in ImageNet classification on VGG16, and 1.8x speedup with 0.025 mAP degradation in PASCAL VOC object detection with Faster-RCNN. ViP also reduces mobile GPU and CPU energy consumption by up to 55% and 70%, respectively. As a complementary method to existing acceleration approaches, ViP achieves 1.9x speedup on ThiNet leading to a combined speedup of 5.23x on VGG16. Furthermore, ViP provides a knob for machine learning practitioners to generate a set of CNN models with varying trade-offs between system speed/energy consumption and accuracy to better accommodate the requirements of their tasks. Code is available at https://github.com/cmu-enyac/VirtualPooling.

****

Learn a Global Appearance Semi-Supervisedly for Synthesizing Person Images

Zhipeng Ge, Fei Chen, Yu Zhou, Yao Yu, Sidan Du; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1190-1199

We present a novel approach for person images synthesis in this paper, that can generate person images in arbitrary poses, shapes and views. Unlike existing methods just using keypoints' locations in heatmaps format, we propose to render SMPL model to UV maps, which can provide human structural information about poses and shapes. Thus, by varying the parameters of poses, shapes and camera in SMPL model, we can generate different person images with various attributions in a simple way, while in most cases we can only obtain new shapes of people by computer graphics methods. We train an end to end generative adversarial network with unlabeled data. As our SMPL parameters come from a pretrained model, we call our overall network semi-supervised. Our network keeps a global appearance during the fine-tuning stage of the target person, thus we can get a complete appearance of the target person, rather than the inaccurate appearance caused by inferencing without enough information. Experiments on Human3.6M Dataset and a self-collected dataset demonstrate the excellent effectiveness of our approach on person images synthesis for different applications.

****

ReStGAN: A step towards visually guided shopper experience via text-to-image synthesis

Shiv Surya, Amrith Setlur, Arijit Biswas, Sumit Negi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1200-1208

E-commerce companies like Amazon, Alibaba and Flipkart have an extensive catalogue comprising of billions of products. Matching customer search queries to plausible products is challenging due to the size and diversity of the catalogue. These challenges are compounded in apparel due to the semantic complexity and a large variation of fashion styles, product attributes and colours. Providing aids that can help the customer visualise the styles and colours matching their "search queries" will provide customers with necessary intuition about what can be done next. This helps the customer buy a product with the styles, embellishments and colours of their liking. In this work, we propose a Generative Adversarial Network (GAN) for generating images from text streams like customer search queries. Our GAN learns to incrementally generate possible images complementing the fine-grained style, colour of the apparel in the query. We incorporate a novel colour modelling approach enabling the GAN to render a wide spectrum of colours accurately. We compile a dataset from an e-commerce website to train our model. The proposed approach outperforms the baselines on qualitative and quantitative evaluations.

****

A Multi-Space Approach to Zero-Shot Object Detection

Dikshant Gupta, Aditya Anantharaman, Nehal Mamgain, Sowmya Kamath S, Vineeth N Balasubramanian, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1209-1217

Object detection has been at the forefront for higher level vision tasks such as scene understanding and contextual reasoning. Therefore, solving object detection for a large number of visual categories is paramount. Zero-Shot Object Detection (ZSD) - where training data is not available for some of the target classes - provides semantic scalability to object detection and reduces dependence on large amount of annotations, thus enabling a large number of applications in real-life scenarios. In this paper, we propose a novel multi-space approach to solve ZSD where we combine predictions obtained in two different search spaces. We learn the projection of visual features of proposals to the semantic embedding space and class labels in the semantic embedding space to visual space. We predict similarity scores in the individual spaces and combine them. We present promising results on two datasets, PASCAL VOC and MS COCO. We further discuss the problem of hubness and show that our approach alleviates hubness with a performance superior to previously proposed methods.

****

L*ReLU: Piece-wise Linear Activation Functions for Deep Fine-grained Visual Categorization

Mina Basirat, PETER ROTH; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1218-1227

Deep neural networks paved the way for significant improvements in image visual categorization during the last years. However, even though the tasks are highly varying, differing in complexity and difficulty, existing solutions mostly build on the same architectural decisions. This also applies to the selection of activation functions (AFs), where most approaches build on Rectified Linear Units (ReLUs). In this paper, however, we show that the choice of a proper AF has a significant impact on the classification accuracy, in particular, if fine, subtle details are of relevance. Therefore, we propose to model the absence and the presence of features via the AF by using piece-wise AFs, which we refer to as L*ReLU. In this way, we can ensure the required properties, while still inheriting the benefits in terms of computational efficiency. We demonstrate our approach for the tasks of Fine-grained Visual Categorization (FGVC), running experiments on seven different benchmark datasets. The results do not only demonstrate superior results but also that for different tasks, having different characteristics, diff erent AFs are selected.

****

An Adversarial Domain Adaptation Network for Cross-Domain Fine-Grained Recognition

Yimu Wang, Renjie Song, Xiu-Shen Wei, Lijun Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1228-1236

In this paper, we tackle a valuable yet very challenging visual recognition task, where the instances are within a subordinate category, and the target domain undergoes a shift with the source domain. This task, termed as cross-domain fine-grained recognition, relates closely to many real-life scenarios, e.g., recognizing retail products in storage racks by models trained with images collected in controlled environments. To deal with this problem, we design a new algorithm and propose a corresponding fine-grained domain adaptation dataset. Firstly, we propose a novel end-to-end CNN architecture that integrates two specialized modules: an adversarial module for domain alignment and a self-attention module for fine-grained recognition. The adversarial module is used to handle domain shift by gradually aligning the different domains with domain-level and class-level alignments, and strive to help the classifier learn with domain-invariant features generated by nets. The self-attention module is designed to capture discriminative image regions which are crucial for fine-grained visual recognition. Secondly, we collect a large-scale fine-grained domain adaptation dataset of retail products, which contains 52,011 images of 263 classes from 3 domains. Thirdly, we validate the effectiveness of our method on three datasets, showing that the proposed method can yield significant improvements over baseline methods on fine-grained datasets. Besides, we also evaluate the effectiveness of the self-attention module by performing visualization, which can capture the discriminative image regions in both source and target domains.

****

Generative Model with Semantic Embedding and Integrated Classifier for Generalized Zero-Shot Learning

Ayyappa Pambala, Titir Dutta, Soma Biswas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1237-1246

Generative models have achieved impressive performance for the generalized zero-shot learning task by learning the mapping from attributes to feature space. In this work, we propose to derive semantic inferences from images and use them for the generation, which enables us to capture the bidirectional information i.e., visual to semantic and semantic to visual spaces. Specifically, we propose a Semantic Embedding module which not only gives image specific semantic information to the generative model for generation of better features, but also makes sure that the generated features can be mapped to the correct semantic space. We also propose an Integrated Classifier, which is trained along with the generator. This module not only eliminates the requirement of additional classifier for new object categories which is required by the existing generative approaches, but also facilitates the generation of more discriminative and useful features. This approach can be used seamlessly for the task of few-shot learning. Extensive experiments on four benchmark datasets, namely, CUB, SUN, AWA1, AWA2 for both zero-shot learning and few-shot setting show the effectiveness of the proposed approach.

****

ELoPE: Fine-Grained Visual Classification with Efficient Localization, Pooling and Embedding

Harald Hanselmann, Hermann Ney; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1247-1256

The task of fine-grained visual classification (FGVC) deals with classification problems that display a small inter-class variance such as distinguishing between different bird species or car models. State-of-the-art approaches typically tackle this problem by integrating an elaborate attention mechanism or (part-) localization method into a standard convolutional neural network (CNN). Also in this work the aim is to enhance the performance of a backbone CNN such as ResNet by including three efficient and lightweight components specifically designed for FGVC. This is achieved by using global k-max pooling, a discriminative embedding layer trained by optimizing class means and an efficient localization module that estimates bounding boxes using only class labels for training. The resulting model achieves state-of-the-art recognition accuracies on multiple FGVC benchmark datasets.

****

Scale Match for Tiny Person Detection

Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, Zhenjun Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1257-1265

Visual object detection has achieved unprecedented advance with the rise of deep convolutional neural networks.However, detecting tiny objects (for example tiny persons less than 20 pixels) in large-scale images remains challenging. The extremely small objects raise a grand challenge about feature representation while the massive and complex backgrounds aggregates the risk of false detections. In this paper, we introduce a new benchmark, referred to as TinyPerson, opening up a promising direction for tiny object detection in a long distance and with massive back-grounds. We experimentally find that the scale mismatch be-tween the dataset for network pretraining and the dataset for detector learning could deteriorate the feature representation and the detectors. Accordingly, we propose a simple yet effective Scale Match approach to align the object scales between the two datasets for favorable tiny-object representation. Experiments show the significant performance gain of our proposed approach over state-of-the-art detectors, and the challenging aspects of TinyPerson related to real-world scenarios. The TinyPerson benchmark and the code for our approach will be publicly available.

****

ScaIL: Classifier Weights Scaling for Class Incremental Learning

Eden Belouadah, Adrian Popescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1266-1275

Incremental learning is useful if an AI agent needs to integrate data from a stream. The problem is non trivial if the agent runs on a limited computational budget and has a bounded memory of past data. In a deep learning approach, the constant computational budget requires the use of a fixed architecture for all incremental states. The bounded memory generates imbalance in favor of new classes and a prediction bias toward them appears. This bias is commonly countered by introducing a data balancing step in addition to the basic network training. We depart from this approach and propose simple but efficient scaling of past classifiers' weights to make them more comparable to those of new classes. Scaling exploits incremental state statistics and is applied to the classifiers learned in the initial state of classes to profit from all their available data. We also question the utility of the widely used distillation loss component of incremental learning algorithms by comparing it to vanilla fine tuning in presence of a bounded memory. Evaluation is done against competitive baselines using four public datasets. Results show that the classifier weights scaling and the removal of the distillation are both beneficial.

****

Extracting identifying contours for African elephants and humpback whales using a learned appearance model

Hendrik Weideman, Chuck Stewart, Jason Parham, Jason Holmberg, Kiirsten Flynn, John Calambokidis, D. Barry Paul, Anka Bedetti, Michelle Henley, Frank Pope, Jerenimo Lepirei; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1276-1285

This paper addresses the problem of identifying individual animals in images based on extracting and matching contours, focusing in particular on the trailing edges of humpback whale flukes and the outline of the ears of African savanna elephants. A coarse-grained FCNN is learned to isolate the contour in an image, and a fine-grained FCNN is learned to provide more precise boundary information. The latter is trained by generating synthetic boundaries from coarse, easily-extra cted training data, avoiding tedious manual effort. An A* algorithm extracts the final contour, which is converted to set of digital curvature descriptors and matched against a database of descriptors using local-naive Bayes nearest neighbors. We show that using the learned fine-grained FCNN produces more accurate contours than using image gradients for fine localization, especially for elephant ears where the boundaries are primarily texture. Matching using contours extracted using the fine-grained FCNN improves top-1 accuracy from 80% to 85% for flukes and 78% to 84% for ears.

****

Anchor Box Optimization for Object Detection

Yuanyi Zhong, Jianfeng Wang, Jian Peng, Lei Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1286-1294

In this paper, we propose a general approach to optimize anchor boxes for object detection. Nowadays, anchor boxes are widely adopted in state-of-the-art detection frameworks. However, these frameworks usually pre-define anchor box shapes in heuristic ways and fix the sizes during training. To improve the accuracy and reduce the effort of designing anchor boxes, we propose to dynamically learn the anchor shapes, which allows the anchors to automatically adapt to the data distribution and the network learning capability. The learning approach can be easily implemented with stochastic gradient descent and can be plugged into any anchor box-based detection framework. The extra training cost is almost negligible and it has no impact on the inference time or memory cost. Exhaustive experiments demonstrate that the proposed anchor optimization method consistently achieves significant improvement (>1% mAP absolute gain) over the baseline methods on several benchmark datasets including Pascal VOC 07+12, MS COCO and Brainwash. Meanwhile, the robustness is also verified towards different anchor initialization methods and the number of anchor shapes, which greatly simplifies the problem of anchor box design.

****

GAR: Graph Assisted Reasoning for Object Detection

Zheng Li, Xiaocong Du, Yu Cao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1295-1304

It is well believed that object-object relations and object-scene relations inherently improve the accuracy of object detection. However, the way to efficiently model relations remains a problem. Graph Convolutional Network (GCN), an effective method to handle structured data with relations, inspires us to leverage graphs in modeling relations for objection detection tasks. In this work, we propose a novel approach, Graph Assisted Reasoning (GAR), to utilize a heterogeneous graph in modeling object-object relations and object-scene relations. GAR fuses the features from neighboring object nodes as well as scene nodes and produces better recognition than that produced from individual object nodes. Moreover, compared to previous approaches using Recurrent Neural Network (RNN), the light-weight and low-coupling architecture of GAR further facilitates its integration into the object detection module. Comprehensive experiments on PASCAL VOC and MS COCO datasets demonstrate the efficacy of GAR.

****

Improving Object Detection with Inverted Attention

zeyi huang, Wei Ke, Dong Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1305-1313

Improving object detectors against occlusion, blur and noise is a critical step to deploy detectors in real applications. Since it is not possible to exhaust all image defects and occlusions through data collection, many researchers seek to generate occluded samples. The generated hard samples are either images or feature maps with coarse patches dropped out in the spatial dimensions. Significant overheads are required in generating hard samples and/or estimating drop-out patches using extra network branches. In this paper, we improve object detectors using a highly efficient and fine-grain mechanism called Inverted Attention (IA). Different from the original detector network that only focuses on the dominant part of objects, the detector network with IA iteratively inverts attention on fe

ature maps which push the detector to discover new discriminative clues and puts more attention on complementary object parts, feature channels and even context. Our approach (1) operates along both the spatial and channels dimensions of the feature maps; (2) requires no extra training on hard samples, no extra network parameters for attention estimation, and no testing overheads. Experiments show that our approach consistently improved state-of-the-art detectors on benchmark databases.
****

Instance Segmentation of Benthic Scale Worms at a Hydrothermal Site

Bhuvan Malladihalli Shashidhara, Mitchell Scott, Aaron Marburg; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1314-1323

Subsea hydrothermal vents, typically existing at water depths below natural light penetration, contain diverse and unique macrofaunal environments. Traditionally, long-term ecological observation has been difficult as the extreme depth, temperature and pressure make in situ video surveys challenging. However, the introduction of subsea cabled arrays has allowed for the long time series collection of high definition imagery from these vents. To study the benthic hydrothermal vent environment, we propose an inference pipeline consisting of a U-Net followed by VGG-16 CNN to perform instance segmentation of scale worms, a specific macrofaunal family. The developed pipeline exhibits an average precision (AP) of 0.671 AP@[0.5], despite the difficult camouflaged imagery and low training data inputs. We further explore full pipeline training data requirements, as the dynamic scene in question requires the pipeline to be re-trained on an approximately monthly basis for effective segmentation. We find that the VGG-16 CNN portion of the pipeline is typically more sensitive to training data variation than the U-Net portion.
****

Multi-Scale Adversarial Cross-Domain Detection with Robust Discriminative Learning

YoungSun Pan, Andy J Ma, Yuan Gao, Jinpeng Wang, YiQi Lin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1324-1332

Domain shift practically exists in almost all computer vision tasks including object detection, caused by which the performance drops evidently. Most existing methods for domain adaptation are specially designed for classification. For object detection, existing methods separate domain shift into image-level shift and instance-level shift and align image-level feature and instance-level feature respectively. However, we find that there are two problems which remain unsolved yet. First, the scale of objects is not the same even in an image. Second, negative transfer can affect model performance if not handled properly. We improve the performance of cross-domain detection from three perspectives: 1) using multiple dilated convolution kernels with different dilation rate to reduce the image-level domain discrepancy; 2) removing images or instances with low transferability to weaken the influence of negative transfer; 3) diversifying distributions by keeping instances' feature away from each other, and then pull them closer to the center of each category, so that make source samples distribution more dispersed and more robust for cross-domain detection. We test our model with Cityscapes, Foggy Cityscape and SIM 10K datasets, experimental results show that our method outperforms the state-of-the-art for object detection under the setting of unsupervised domain adaptation (UDA).
****

Combining Compositional Models and Deep Networks For Robust Object Classification under Occlusion

Adam Kortylewski, Qing Liu, Huiyu Wang, Zhishuai Zhang, Alan Yuille; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1333-1341

Deep convolutional neural networks (DCNNs) are powerful models that yield impressive results at object classification. However, recent work has shown that they do not generalize well to partially occluded objects and to mask attacks. In con trast to DCNNs, compositional models are robust to partial occlusion, however, they are not as discriminative as deep models. In this work, we combine DCNNs and compositional object models to retain the best of both approaches: a discriminative model that is robust to partial occlusion and mask attacks. Our model is learned in two steps. First, a standard DCNN is trained for image classification. Subsequently, we cluster the DCNN features into dictionaries. We show that the dictionary components resemble object part detectors and learn the spatial distribution of parts for each object class. We propose mixtures of compositional models to account for large changes in the spatial activation patterns (e.g. due to changes in the 3D pose of an object). At runtime, an image is first classified by the DCNN in a feedforward manner. The prediction uncertainty is used to detect partially occluded objects, which in turn are classified by the compositional model. Our experimental results demonstrate that combining compositional models and DCNNs resolves a fundamental problem of current deep learning approaches to computer vision: The combined model recognizes occluded objects, even when it has not been exposed to occluded objects during training, while at the same time maintaining high discriminative performance for non-occluded objects.

****

Robust estimation of local affine maps and its applications to image matching

Mariano RODRIGUEZ, Gabriele Facciolo, Rafael Grompone von Gioi, Pablo Muse, Julie Delon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1342-1351

The classic approach to image matching consists in the detection, description and matching of keypoints. This defines a zero-order approximation of the mapping between two images, determined by corresponding point coordinates. But the patches around keypoints typically contain more information, which may be exploited to obtain a first-order approximation of the mapping, incorporating local affine maps between corresponding keypoints. In this work, we propose a LOCal Affine Transform Estimator (LOCATE) method based on neural networks. We show that LOCATE drastically improves the accuracy of local geometry estimation by tracking inverse maps. A second contribution on guided matching and refinement is presented. The novelty here consists in the use of LOCATE to propose new SIFT-keypoint correspondences with precise locations, orientations and scales. Our experiments show that the precision gain provided by LOCATE does play an important role in applications such as guided matching. The third contribution of this paper consists in a modification to the RANSAC algorithm, that use LOCATE to improve the homography estimation between a pair of images. These approaches outperform RANSAC for different choices of image descriptors and image datasets, and permit to increase the probability of success in identifying image pairs in challenging matching databases.

****

Multi-way Encoding for Robustness

Donghyun Kim, Sarah Bargal, Jianming Zhang, Stan Sclaroff; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1352-1360

Deep models are state-of-the-art for many computer vision tasks including image classification and object detection. However, it has been shown that deep models are vulnerable to adversarial examples. We highlight how one-hot encoding directly contributes to this vulnerability and propose breaking away from this widely-used, but highly-vulnerable mapping. We demonstrate that by leveraging a different output encoding, multi-way encoding, we decorrelate source and target models, making target models more secure. Our approach makes it more difficult for adversaries to find useful gradients for generating adversarial attacks. We present robustness for black-box and white-box attacks on four benchmark datasets: MNIST, CIFAR-10, CIFAR-100, and SVHN. The strength of our approach is also presented in the form of an attack for model watermarking, raising challenges in detecting stolen models.

****

Robust Face Detection via Learning Small Faces on Hard Images

Zhishuai Zhang, Wei Shen, Siyuan Qiao, Yan Wang, Bo Wang, Alan Yuille; Proc eedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1361-1370

Recent anchor-based deep face detectors have achieved promising performance, but they are still struggling to detect hard faces, such as small, blurred and partially occluded faces. One reason is that they treat all images and faces equally, and ignore the imbalance between easy images and hard images; however large amounts of training images only contain easy faces, which are less helpful to learn robust detectors for hard faces. In this paper, we propose that the robustness of a face detector against hard faces can be improved by learning small faces on hard images. Our intuitions are (1) hard images are the images which contain at least one hard face, thus they facilitate training robust face detectors; (2) most hard faces are small faces and other types of hard faces can be easily shrunk to small faces. To this end, we build an anchor-based deep face detector, which only outputs a single high-resolution feature map with small anchors, to specifically learn small faces and train it by a novel hard image mining strategy which automatically adjusts training weights on images according to their difficulties. Extensive experiments have been conducted on WIDER FACE, FDDB, Pascal Faces, and AFW datasets and our method achieves APs of 95.7, 94.9 and 89.7 on easy, medium and hard WIDER FACE val dataset respectively, which verify the effectiveness of our methods, especially on detecting hard faces. Our detector is also lightweight and enjoys a fast inference speed. Code and model are available at https://github.com/bairdzhang/smallhardface.

****

Deep Learning on Small Datasets without Pre-Training using Cosine Loss

Bjorn Barz, Joachim Denzler; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1371-1380

Two things seem to be indisputable in the contemporary deep learning discourse: 1. The categorical cross-entropy loss after softmax activation is the method of choice for classification. 2. Training a CNN classifier from scratch on small datasets does not work well. In contrast to this, we show that the cosine loss function provides substantially better performance than cross-entropy on datasets with only a handful of samples per class. For example, the accuracy achieved on the CUB-200-2011 dataset without pre-training is by 30% higher than with the cross-entropy loss. Further experiments on other popular datasets confirm our findings. Moreover, we demonstrate that integrating prior knowledge in the form of class hierarchies is straightforward with the cosine loss and improves classification performance further.

****

A Novel Self-Supervised Re-labeling Approach for Training with Noisy Labels

Devraj Mandal, Shrisha Bharadwaj, Soma Biswas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1381-1390

The major driving force behind the immense success of deep learning models is the availability of large datasets along with their clean labels. This is very difficult to obtain and thus has motivated research on training deep neural networks in the presence of label noise. In this work, we build upon the seminal work in this area, Co-teaching and propose a simple, yet efficient approach termed mCT - S2R (modified co-teaching with self-supervision and re-labeling) for this task. Firstly, to deal with significant amount of noise in the labels, we propose to use self- supervision to generate robust features without using any labels. Furthermore, using a parallel network architecture, an estimate of the clean labeled portion of the data is obtained. Finally, using this data, a portion of the estimated noisy labeled portion is re-labeled, before resuming the network training with the augmented data. Extensive experiments on three standard datasets show the effectiveness of the proposed framework.

****

Crowded Human Detection via an Anchor-pair Network

Jinguo Zhu, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, shenghao zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1391-1399

This paper presents an anchor-pair network for crowded human detection, which ca n overcome and solve the difficulties caused by occlusion in crowded scenes. Specifically, we use a function-aware network structure to extract more distinctive and discriminative features for head and full-body respectively, and then a CNN module is also exploited to fuse the features by learning the correlations between head and full-body to reduce crowd errors. Meanwhile, a novel paired form for anchors, denoted as anchor-pair, is proposed to estimate the head regions and full-body regions simultaneously. Furthermore, a new ingenious JointNMS is introduced to perform on the detected head and full-body box pairs, which produces significant performance improvement in heavily occluded scenarios at tiny computational cost. Our anchor-pair network achieves a state-of-the-art result on the CrowdHuman dataset which reduces the MR 2 to 55.43%, achieving 11.59% relative improvement over our dataset baseline.

****

Multi-class Novelty Detection Using Mix-up Technique

Supritam Bhattacharjee, Devraj Mandal, Soma Biswas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1400-1409

Multi-class novelty detection is increasingly becoming an important area of research due to the continuous increase in the number of object categories. It tries to answer the pertinent question: given a test sample, should we even try to classify it? We propose a novel solution using the concept of mix-up technique for novelty detection, termed as Segregation Network. During training, a pair of examples are selected from the training data and an interpolated data point using their convex combination is constructed. We develop a suitable loss function to train our model to predict its constituent classes. During testing, each input query is combined with the known class prototypes to generate mixed samples which are then passed through the trained network. Our model which is trained to reveal the constituent classes can then be used to determine whether the sample is novel or not. The intuition is that if a query comes from a known class and is mixed with the set of known class prototypes, then the prediction of the trained model for the correct class should be high. In contrast, for a query from a novel class, the predictions for all the known classes should be low. The proposed model is trained using only the available known class data and does not need access to any auxiliary dataset or attributes. Extensive experiments on two benchmark datasets, namely Caltech 256 and Stanford Dogs and comparisons with the state-of-the-art algorithms justifies the usefulness of our approach.

****

Analysis and a Solution of Momentarily Missed Detection for Anchor-based Object Detectors

Yusuke Hosoya, Masanori Suganuma, Takayuki Okatani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1410-1418

The employment of convolutional neural networks has led to significant performance improvement on the task of object detection. However, when applying existing detectors to continuous frames in a video, we often encounter momentary miss-detection of objects, that is, objects are undetected exceptionally at a few frames, although they are correctly detected at all other frames. In this paper, we analyze the mechanism of how such miss-detection occurs. For the most popular class of detectors that are based on anchor boxes, we show the followings: i) besides apparent causes such as motion blur, occlusions, background clutters, etc., the majority of remaining miss-detection can be explained by an improper behavior of the detectors at boundaries of the anchor boxes; and ii) this can be rectified by improving the way of choosing positive samples from candidate anchor boxes when training the detectors.

****

DATNet: Dense Auxiliary Tasks for Object Detection

Alex Levinshtein, Alborz Rezazadeh Sereshkeh, Konstantinos Derpanis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1419-1427

Beginning with R-CNN, there has been a rapid advancement in two-stage object det ection approaches. While two-stage approaches remain the state-of-the-art in object detection, anchor-free single-stage methods have been gaining momentum. We believe that the strength of the former is in their region of interest (ROI) pooling stage, while the latter simplifies the learning problem by converting object detection into dense per-pixel prediction tasks. In this paper, we propose to combine the strengths of each approach in a new architecture. In particular, we first define several auxiliary tasks related to object detection and generate dense per-pixel predictions using a shared feature extraction backbone. As a consequence of this architecture, the shared backbone is trained using both the standard object detection losses and these per-pixel ones. Moreover, by combining the features from dense predictions with those from the backbone, we realize a more discriminative representation for subsequent downstream processing. In addition, we feed the fused features into a novel multi-scale ROI pooling layer, followed by per-ROI predictions. We refer to our architecture as the Dense Auxiliary Tasks Network (DATNet). We present an extensive set of evaluations of our method on the Pascal VOC and COCO datasets and show considerable accuracy improvements over comparable baselines.

****

Active Learning for Imbalanced Datasets

Umang Aggarwal, Adrian Popescu, Celine Hudelot; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1428-1437

Active learning increases the effectiveness of labeling when only subsets of unlabeled datasets can be processed manually. To our knowledge, existing algorithms are designed under the assumption that datasets are balanced. However, many real-life datasets are actually imbalanced and we propose two adaptations of active learning to tackle imbalance. First, we modify acquisition functions to select samples by taking advantage of a deep model pretrained on a source domain. Second, we introduce a balancing step in the acquisition process to reduce the imbalance of the labeled subset. Evaluation is done with four imbalanced datasets using existing active learning methods and their modifications introduced here. Results show that our adaptations are useful as long as knowledge from the source domain is transferable to target domains.

****

Animal Detection in Man-made Environments

Abhineet Singh, Marcin Pietrasik, Gabriell Natha, Nehla Ghouaiel, Ken Brizel, Nilanjan Ray; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1438-1449

Automatic detection of animals that have strayed into human inhabited areas has important security and road safety applications. This paper attempts to solve this problem using deep learning techniques from a variety of computer vision fields including object detection, tracking, segmentation and edge detection. Several interesting insights into transfer learning are elicited while adapting models trained on benchmark datasets for real world deployment. Empirical evidence is presented to demonstrate the inability of detectors to generalize from training images of animals in their natural habitats to deployment scenarios of man-made environments. A solution is also proposed using semi-automated synthetic data generation for domain specific training. Code and data used in the experiments are made available to facilitate further work in this domain.

****

CookGAN: Meal Image Synthesis from Ingredients

Fangda Han, Ricardo Guerrero, Vladimir Pavlovic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1450-1458

In this work we propose a new computational framework, based on generative deep models, for synthesis of photo-realistic food meal images from textual list of its ingredients. Previous works on synthesis of images from text typically rely on pre-trained text models to extract text features, followed by generative neural networks (GAN) aimed to generate realistic images conditioned on the text features. These works mainly focus on generating spatially compact and well-defined categories of objects, such as birds or flowers, but meal images are significantly more complex, consisting of multiple ingredients whose appearance and spatial qualities are further modified by cooking methods. To generate real-like meal images from ingredients, we propose Cook Generative Adversarial Networks (CookGAN), CookGAN first builds an attention-based ingredients-image association model, which is then used to condition a generative neural network tasked with synthesizing meal images. Furthermore, a cycle-consistent constraint is added to further improve image quality and control appearance. Experiments show our model is able to generate meal images corresponding to the ingredients.

****

Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison

DONGXU LI, Cristian Rodriguez, Xin Yu, HONGDONG LI; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1459-1469

Vision-based sign language recognition aims at helping the hearing-impaired people to communicate with others. However, most existing sign language datasets are limited to a small number of words. This makes the migration of recognition systems to real-life scenarios difficult. Due to the limited vocabulary size, models learned from those datasets cannot be applied in practice. In this paper, we introduce a new large-scale Word-Level American Sign Language (WLASL) video dataset, containing more than 2000 words performed by over 100 signers. This dataset will be made publicly available to the research community. To our knowledge, it is by far the largest public ASL dataset to facilitate word-level sign recognition research.   Based on this new large-scale dataset, we are able to experiment with several deep learning methods for word-level sign recognition and evaluate their performances in large scale scenarios. Specifically we implement and compare two different models,i.e., (i) holistic visual appearance based approach, and (ii) 2D human pose based approach. Both models are valuable baselines that will benefit the community for method benchmarking. Moreover, we also propose a novel pose-based temporal graph convolution networks (Pose-TGCN) that model spatial and temporal dependencies in human pose trajectories simultaneously, which has further boosted the performance of the pose-based method. Our results show that pose-based and appearance-based models achieve comparable performances up to 62.63% at top-10 accuracy on 2,000 words/glosses, demonstrating the validity and challenges of our dataset. Our dataset and baseline deep mod- els are available at https://dxli94.github.io/WLASL/.

****

Exploring Hate Speech Detection in Multimodal Publications

Raul Gomez, Jaume Gibert, Lluis Gomez, Dimosthenis Karatzas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1470-1478

In this work we target the problem of hate speech detection in multimodal publications formed by a text and an image. We gather and annotate a large scale dataset from Twitter, MMHS150K, and propose different models that jointly analyze textual and visual information for hate speech detection, comparing them with unimodal detection. We provide quantitative and qualitative results and analyze the challenges of the proposed task. We find that, even though images are useful for the hate speech detection task, current multimodal models cannot outperform models analyzing only text. We discuss why and open the field and the dataset for further research.

****

Bridged Variational Autoencoders for Joint Modeling of Images and Attributes

Ravindra Yadav, Ashish Sardana, Vinay Namboodiri, Rajesh M Hegde; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1479-1487

Generative models have recently shown the ability to realistically generate data and model the distribution accurately. However, joint modeling of an image with the attribute that it is labeled with requires learning a cross modal correspondence between images and the attribute data. Though the information present in the images and attributes possess completely different statistical properties altogether, there exists an inherent correspondence that is challenging to capture.

Various models have aimed at capturing this correspondence either through joint modeling of a variational autoencoder or through separate encoder networks that are then concatenated. We present an alternative by proposing a bridged variational autoencoder that allows for learning cross-modal correspondence by incorporating cross-modal hallucination losses in the latent space. In comparison to the existing methods, we have found that by incorporating this information into the network we not only obtain better generation results, but also obtain very distinctive latent embeddings thereby increasing the accuracy of cross-modal generated results. We validate the proposed method through comparison with state of the art methods and benchmarking on standard datasets.

****

Differentiable Scene Graphs

Moshiko Raboh, Roei Herzig, Jonathan Berant, Gal Chechik, Amir Globerson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1488-1497

Reasoning about complex visual scenes involves perception of entities and their relations. Scene Graphs (SGs) provide a natural representation for reasoning tasks, by assigning labels to both entities (nodes) and relations (edges). Reasoning systems based on SGs are typically trained in a two-step procedure: first, a model is trained to predict SGs from images, and next a separate model is trained to reason based on the predicted SGs. However, it would seem preferable to train such systems in an end-to-end manner. The challenge, which we address here is that scene-graph representations are non-differentiable and therefore it isn't clear how to use them as intermediate components. Here we propose Differentiable Scene Graphs (DSGs), an image representation that is amenable to differentiable end-to-end optimization, and requires supervision only from the downstream tasks. DSGs provide a dense representation for all regions and pairs of regions, and do not spend modelling capacity on regions of the image that do not contain objects or relations of interest. We evaluate our model on the challenging task of identifying referring relationships (RR) in three benchmark datasets: Visual Genome, VRD and CLEVR. Using DSGs as an intermediate representation leads to new state-of-the-art performance. The full code is available at https://github.com/shikorab/DSG.

****

Answering Questions about Data Visualizations using Efficient Bimodal Fusion

Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, Christopher Kanan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1498-1507

Chart question answering (CQA) is a newly proposed visual question answering (VQA) task where an algorithm must answer questions about data visualizations, e.g. bar charts, pie charts, and line graphs. CQA requires capabilities that natural-image VQA algorithms lack: fine-grained measurements, optical character recognition, and handling out-of-vocabulary words in both questions and answers. Without modifications, state-of-the-art VQA algorithms perform poorly on this task. Here, we propose a novel CQA algorithm called parallel recurrent fusion of image and language (PReFIL). PReFIL first learns bimodal embeddings by fusing question and image features and then intelligently aggregates these learned embeddings to answer the given question. Despite its simplicity, PReFIL greatly surpasses state-of-the art systems and human baselines on both the FigureQA and DVQA datasets. Additionally, we demonstrate that PReFIL can be used to reconstruct tables by asking a series of questions about a chart.

****

Cross-modal Scene Graph Matching for Relationship-aware Image-Text Retrieval

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1508-1517

Image-text retrieval of natural scenes has been a popular research topic. Since image and text are heterogeneous cross-modal data, one of the key challenges is how to learn comprehensive yet unified representations to express the multi-modal data. A natural scene image mainly involves two kinds of visual concepts, obje

cts and their relationships, which are equally essential to image-text retrieval. Therefore, a good representation should account for both of them. In the light of recent success of scene graph in many CV and NLP tasks for describing complex natural scenes, we propose to represent image and text with two kinds of scene graphs: visual scene graph (VSG) and textual scene graph (TSG), each of which is exploited to jointly characterize objects and relationships in the corresponding modality. The image-text retrieval task is then naturally formulated as cross-modal scene graph matching. Specifically, we design two particular scene graph encoders in our model for VSG and TSG, which can refine the representation of each node on the graph by aggregating neighborhood information. As a result, both object-level and relationship-level cross-modal features can be obtained, which favorably enables us to evaluate the similarity of image and text in the two levels in a more plausible way. We achieve state-of-the-art results on Flickr30k and MS COCO, which verifies the advantages of our graph matching based approach for image-text retrieval.
****
MHSAN: Multi-Head Self-Attention Network for Visual Semantic Embedding
Geondo Park, Chihye Han, Wonjun Yoon, Daeshik Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1518-1526
Visual-semantic embedding enables various tasks such as image-text retrieval, image captioning, and visual question answering. The key to successful visual-semantic embedding is to express visual and textual data properly by accounting for their intricate relationship. While previous studies have achieved much advance by encoding the visual and textual data into a joint space where similar concepts are closely located, they often represent data by a single vector ignoring the presence of multiple important components in an image or text. Thus, in addition to the joint embedding space, we propose a novel multi-head self-attention network to capture various components of visual and textual data by attending to important parts in data. Our approach achieves the new state-of-the-art results in image-text retrieval tasks on MS-COCO and Flicker30K datasets. Through the visualization of the attention maps that capture distinct semantic components at multiple positions in the image and the text, we demonstrate that our method achieves an effective and interpretable visual-semantic joint space.
****
PlotQA: Reasoning over Scientific Plots
Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, Pratyush Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1527-1536
Existing synthetic datasets (FigureQA, DVQA) for reasoning over plots do not contain variability in data labels, real-valued data, or complex reasoning questions. Consequently, proposed models for these datasets do not fully address the challenge of reasoning over plots. In particular, they assume that the answer comes either from a small fixed size vocabulary or from a bounding box within the image. However, in practice, this is an unrealistic assumption because many questions require reasoning and thus have real-valued answers which appear neither in a small fixed size vocabulary nor in the image. In this work, we aim to bridge this gap between existing datasets and real-world plots. Specifically, we propose PlotQA with 28.9 million question-answer pairs over 224,377 plots on data from real-world sources and questions based on crowd-sourced question templates. Further, 80.76% of the out-of-vocabulary (OOV) questions in PlotQA have answers that are not in a fixed vocabulary. Analysis of existing models on PlotQA reveals that they cannot deal with OOV questions: their overall accuracy on our dataset is in single digits. This is not surprising given that these models were not designed for such questions. As a step towards a more holistic model which can address fixed vocabulary as well as OOV questions, we propose a hybrid approach: Specific questions are answered by choosing the answer from a fixed vocabulary or by extracting it from a predicted bounding box in the plot, while other questions are answered with a table question-answering engine which is fed with a structured table generated by detecting visual elements from the image. On the existin

g DVQA dataset, our model has an accuracy of 58%, significantly improving on the highest reported accuracy of 46%. On PlotQA, our model has an accuracy of 22.52%, which is significantly better than state of the art models.

****

Figure Captioning with Relation Maps for Reasoning

Charles Chen, Ruiyi Zhang, Eunyee Koh, Sungchul Kim, Scott Cohen, Ryan Rossi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1537-1545

Figures, such as line plots, pie charts, bar charts, are widely used to convey important information in a concise format. In this work, we investigate the problem of figure caption generation where the goal is to automatically generate a natural language description for a given figure. While natural image captioning has been studied extensively, figure captioning has received relatively little attention and remains a challenging problem. A successful solution to this task has many potential applications, such as: 1) adding captions to the output of a visualization tool; 2) summarizing documents with a number of figures with or without proper captions; 3) improving user experience by allowing figure content to be accessible to those with visual impairment. To solve this problem, we collect a dataset FigCAP for testing the capability of generating captions, and propose a captioning framework with novel attention models. In order to solve the exposure bias issue, we further train the captioning model with sequence-level policy based on reinforcement learning, which directly optimizes evaluation metrics. Extensive experiments show that our proposed models outperform strong image captioning baselines, thus demonstrating a significant potential for automatic generating captions for figures.

****

Rotation-invariant Mixed Graphical Model Network for 2D Hand Pose Estimation

Deying Kong, Haoyu Ma, Yifei Chen, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1546-1555

In this paper, we propose a new architecture named Rotation-invariant Mixed Graphical Model Network (R-MGMN) to solve the problem of 2D hand pose estimation from a monocular RGB image. By integrating a rotation net, the R-MGMN is invariant to rotations of the hand in the image. It also has a pool of graphical models, from which a combination of graphical models could be selected, conditioning on the input image. Belief propagation is performed on each graphical model separately, generating a set of marginal distributions, which are taken as the confidence maps of hand keypoint positions. Final confidence maps are obtained by aggregating these confidence maps together. We evaluate the R-MGMN on two public hand pose datasets. Experiment results show our model outperforms the state-of-the-art algorithm which is widely used in 2D hand pose estimation by a noticeable margin.

****

BERT representations for Video Question Answering

Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, Haruo Takemura; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1556-1565

Visual question answering (VQA) aims at answering questions about the visual content of an image or a video. Currently, most work on VQA is focused on image-based question answering, and less attention has been paid into answering questions about videos. However, VQA in video presents some unique challenges that are worth studying: it not only requires to model a sequence of visual features over time, but often it also needs to reason about associated subtitles. In this work, we propose to use BERT, a sequential modelling technique based on Transformers, to encode the complex semantics from video clips. Our proposed model jointly captures the visual and language information of a video scene by encoding not only the subtitles but also a sequence of visual concepts with a pre-trained language-based Transformer. In our experiments, we exhaustively study the performance of our model by taking different input arrangements, showing outstanding improvements when compared against previous work on two well-known video VQA datasets: TVQA and Pororo.

****

Deep Bayesian Network for Visual Question Generation

Badri Patro, Vinod Kurmi, Sandeep Kumar, Vinay Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1566-1576

Generating natural questions from an image is a semantic task that requires using vision and language modalities to learn multimodal representations. Images can have multiple visual and language cues such as places, captions, and tags. In this paper, we propose a principled deep Bayesian learning framework that combines these cues to produce natural questions. We observe that with the addition of more cues and by minimizing uncertainty in the among cues, the Bayesian network becomes more confident. We propose a Minimizing Uncertainty of Mixture of Cues (MUMC), that minimizes uncertainty present in a mixture of cues experts for generating probabilistic questions. This is a Bayesian framework and the results show a remarkable similarity to natural questions as validated by a human study. We observe that with the addition of more cues and by minimizing uncertainty among the cues, the Bayesian framework becomes more confident. Ablation studies of our model indicate that a subset of cues is inferior at this task and hence the principled fusion of cues is preferred. Further, we observe that the proposed approach substantially improves over state-of-the-art benchmarks on the quantitative metrics (BLEU-n, METEOR, ROUGE, and CIDEr). Here we provide project link for Deep Bayesian VQG https://delta-lab-iitk.github.io/BVQG/.

****

Robust Explanations for Visual Question Answering

Badri Patro, Shivansh Patel, Vinay Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1577-1586

In this paper, we propose a method to obtain robust explanations for visual question answering(VQA) that correlate well with the answers. Our model explains the answers obtained through a VQA model by providing visual and textual explanations. The main challenges that we address are i) Answers and textual explanations obtained by current methods are not well correlated and ii) Current methods for visual explanation do not focus on the right location for explaining the answer. We address both these challenges by using a collaborative correlated module which ensures that even if we do not train for noise based attacks, the enhanced correlation ensures that the right explanation and answer can be generated. We further show that this also aids in improving the generated visual and textual explanations. The use of the correlated module can be thought of as a robust method to verify if the answer and explanations are coherent. We evaluate this model using VQA-X dataset. We observe that the proposed method yields better textual and visual justification that supports the decision. We showcase the robustness of the model against a noise-based perturbation attack using corresponding visual and textual explanations. A detailed empirical analysis is shown.

****

Domain-Specific Semantics Guided Approach to Video Captioning

Hemalatha M, C Chandra Shekhar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1587-1596

In video captioning, the description of a video usually relies on the domain to which the video belongs. Typically, the videos belong to wide range domains such as sports, music, news, cooking, etc. In many cases, a video can be associated with more than one domain. In this paper, we propose an approach to video captioning that uses domain-specific decoders. We build a domain classifier to obtain the estimates of probabilities of a video belonging to different domains. For each video, we identify the top-k domains based on the estimated probabilities. Each video in the training data set is shared in training the domain-specific decoders of top-k labels obtained from the domain classifier. The domain-specific decoders use the domain-specific semantic tags for generating captions. The proposed approach uses the Temporal VLAD for preprocessing the features extracted from 2D-CNN and 3D-CNN features. The preprocessed features provide better feature representation of the videos. The effectiveness of the proposed approach is demonstrated through the results of experimental studies on Microsoft Video Descriptio

n (MSVD) corpus and MSR-VTT dataset.
****

Adapting Style and Content for Attended Text Sequence Recognition

Steven Schwarcz, Alexander Gorban, Xavier Gibert, Dar-Shyang Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1597-1606

In this paper, we address the problem of learning to perform sequential OCR on photos of street name signs in a language for which no labeled data exists. Our approach leverages easily-generated synthetic data and existing labeled data in other languages to achieve reasonable performance on these unlabeled images, through a combination of a novel domain adaptation technique based on gradient reversal and a multi-task learning scheme. In order to accomplish this, we introduce and release two new datasets - Hebrew Street Name Signs (HSNS) and Synthetic Hebrew Street Name Signs (SynHSNS) - while also making use of the existing French Street Name Signs (FSNS) dataset. We demonstrate that by using a synthetic dataset of Hebrew characters and a labeled dataset of French street name signs in natural images, it is possible to achieve a significant improvement on real Hebrew street name sign transcription, where the synthetic Hebrew data and real French data each overlap with different features of the images we wish to transcribe.
****

Visual Question Answering on 360deg Images

Shih-Han Chou, Wei-Lun Chao, Wei-Sheng Lai, Min Sun, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1607-1616

In this work, we introduce VQA 360deg, a novel task of visual question answering on 360deg images. Unlike a normal field-of-view image, a 360deg image captures the entire visual content around the optical center of a camera, demanding more sophisticated spatial understanding and reasoning. To address this problem, we collect the first VQA 360deg dataset, containing around 17,000 real-world image-question-answer triplets for a variety of question types. We then study two different VQA models on VQA 360deg, including one conventional model that takes an equirectangular image (with intrinsic distortion) as input and one dedicated model that first projects a 360deg image onto cubemaps and subsequently aggregates the information from multiple spatial resolutions. We demonstrate that the cubemap-based model with multi-level fusion and attention diffusion performs favorably against other variants and the equirectangular-based models. Nevertheless, the gap between the humans' and machines' performance reveals the need for more advanced VQA 360deg algorithms. We, therefore, expect our dataset and studies to serve as the benchmark for future development in this challenging task. Dataset, code, and pre-trained models are available online.
****

Spatio-Temporal Ranked-Attention Networks for Video Captioning

Anoop Cherian, Jue Wang, Chiori Hori, Tim Marks; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1617-1626

Generating video descriptions automatically is a challenging task that involves a complex interplay between spatio-temporal visual features and language models. Given that videos consist of spatial (frame-level) features and their temporal evolutions, an effective captioning model should be able to attend to these different cues selectively. To this end, we propose a Spatio-Temporal and Temporo-Spatial (STaTS) attention model which, conditioned on the language state, hierarchically combines spatial and temporal attention to videos in two different orders: (i) a spatio-temporal (ST) sub-model, which first attends to regions that have temporal evolution, then temporally pools the features from these regions; and (ii) a temporo-spatial (TS) sub-model, which first decides a single frame to attend to, then applies spatial attention within that frame. We propose a novel LSTM-based temporal ranking function, which we call ranked attention, for the ST model to capture action dynamics. Our entire framework is trained end-to-end. We provide experiments on two benchmark datasets: MSVD and MSR-VTT. Our results demonstrate the synergy between the ST and TS modules, outperforming recent state-of-the-art methods.

****

ULSAM: Ultra-Lightweight Subspace Attention Module for Compact Convolutional Neural Networks

Rajat Saini, Nandan Kumar Jha, Bedanta Das, Sparsh Mittal, C . Krishna Mohan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1627-1636

The capability of the self-attention mechanism to model the long-range dependencies has catapulted its deployment in vision models. Unlike convolution operators, self-attention offers infinite receptive field and enables compute-efficient modeling of global dependencies. However, the existing state-of-the-art attention mechanisms incur high compute and/or parameter overheads, and hence unfit for compact convolutional neural networks (CNNs). In this work, we propose a simple yet effective "Ultra-Lightweight Subspace Attention Mechanism" (ULSAM), which infers different attention maps for each feature map subspace. We argue that leaning separate attention maps for each feature subspace enables multi-scale and multi-frequency feature representation, which is more desirable for fine-grained image classification. Our method of subspace attention is orthogonal and complementary to the existing state-of-the-arts attention mechanisms used in vision models. ULSAM is end-to-end trainable and can be deployed as a plug-and- play module in the pre-existing compact CNNs. Notably, our work is the first attempt that uses a subspace attention mechanism to increase the efficiency of compact CNNs. To show the efficacy of ULSAM, we perform experiments with MobileNet-V1 and MobileNet-V2 as backbone architectures on ImageNet-1K and three fine-grained image classification datasets. We achieve [?]13% and [?]25% reduction in both the FLOPs and parameter counts of MobileNet-V2 with a 0.27% and more than 1% improvement in top-1 accuracy on the ImageNet-1K and fine-grained image classification datasets (respectively). Code and trained models are available at https://github.com/Nandan91/ULSAM .

****

Watch to Listen Clearly: Visual Speech Enhancement Driven Multi-modality Speech Recognition

Bo Xu, Jacob Wang, Cheng Lu, Yandong Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1637-1646

Multi-modality (talking face video and audio) information helps improve speech recognition performance compared to the single modality. In noisy environments, the effect of audio modality is weakened, which further affects the performance of multi-modality speech recognition (MSR). Most of the MSR methods use noisy audio signal as input of the audio modality without any enhancement (filtering the noisy components in the audio signal). In this paper, we propose an audio-enhanced multi-modality speech recognition model. In particular, the proposed model consists of two sub-networks, one is the visual speech enhancement (VE) sub-network and the other is the multi-modality speech recognition (MSR) sub-network. The VE sub-network is able to separate a speaker's voice from background noises when given the corresponding talking face to enhance audio modality. Then the audio modality together with video modality are fed into the MSR sub-network to produce characters. We introduce a pseudo-3D residual network (P3D)-based visual front-end to extract more advantageous visual features. The MSR sub-network is built on top of the Element-wise-Attention Gated Recurrent Unit (EleAtt-GRU) architecture which is more effective than Transformer in long sequences. We demonstrate the effectiveness of audio enhancement for MSR by extensive experiments. The proposed method surpasses the state-of-the-art MSR models on the LRS3-TED dataset and the LRW dataset.

****

Video Object Segmentation-based Visual Servo Control and Object Depth Estimation on a Mobile Robot

Brent Griffin, Victoria Florence, Jason Corso; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1647-1657

To be useful in everyday environments, robots must be able to identify and locate real-world objects. In recent years, video object segmentation has made significant progress on densely separating such objects from background in real and ch

allenging videos. Building off of this progress, this paper addresses the problem of identifying generic objects and locating them in 3D using a mobile robot with an RGB camera. We achieve this by, first, introducing a video object segmentation-based approach to visual servo control and active perception and, second, developing a new Hadamard-Broyden update formulation. Our segmentation-based methods are simple but effective, and our update formulation lets a robot quickly learn the relationship between actuators and visual features without any camera calibration. We validate our approach in experiments by learning a variety of actuator-camera configurations on a mobile HSR robot, which subsequently identifies, locates, and grasps objects from the YCB dataset and tracks people and other dynamic articulated objects in real-time.
****
Robust Feature Tracking in DVS Event Stream using Bezier Mapping
Hochang Seok, Jongwoo Lim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1658-1667
Unlike conventional cameras, event cameras capture the intensity changes at each pixel with very little delay. Such changes are recorded as an event stream with their positions, timestamps, and polarities continuously, thus there is no notion of 'frame' as in conventional cameras. As many applications including 3D pose estimation use 2D trajectories of feature points, it is necessary to detect and track the feature points robustly and accurately in a continuous event stream. In conventional feature tracking algorithms for event streams, the events in fixed time intervals are converted into the event images by stacking the events at their pixel locations, and the features are tracked in the event images. Such simple stacking of events yields blurry event images due to the camera motion, and it can significantly degrade the tracking quality. We propose to align the events in the time intervals along Bezier curves to minimize the misalignment. Since the camera motion is unknown, the Bezier curve is estimated to maximize the variance of the warped event pixels. Instead of the initial patches for tracking, we use the temporally integrated template patches, as it captures rich texture information from accurately aligned events. Extensive experimental evaluations in 2D feature tracking as well as 3D pose estimation show that our method significantly outperforms the conventional approaches.
****
SymGAN: Orientation Estimation without Annotation for Symmetric Objects
Phil Ammirato, Jonathan Tremblay, Ming-Yu Liu, Alexander Berg, Dieter Fox; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1668-1677
Training a computer vision system to predict an object's pose is crucial to improving robotic manipulation, where robots can easily locate and then grasp objects. Some of the key challenges in pose estimation lie in obtaining labeled data and handling objects with symmetries. We explore both these problems of viewpoint estimation (object 3D orientation) by proposing a novel unsupervised training paradigm that only requires a 3D model of the object of interest. We show that we can successfully train an orientation detector, which simply consumes an RGB image, in an adversarial training framework, where the discriminator learns to provide a learning signal to retrieve the object orientation using a black-box non differentiable renderer. In order to overcome this non differentiability, we introduce a randomized sampling method to obtain training gradients. To our knowledge this is the first time an adversarial framework is employed to successfully train a viewpoint detector that can handle symmetric objects.Using this training framework we show state of the art results on 3D orientation prediction on T-LESS, a challenging dataset for texture-less and symmetric objects.
****
QUICKSAL: A small and sparse visual saliency model for efficient inference in resource constrained hardware
Vignesh Ramanathan, Pritesh Dwivedi, Bharath Katabathuni, Anirban Chakraborty, Chetan Singh Thakur; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1678-1688
Visual saliency is an important problem in the field of cognitive science and co

mputer vision with applications such as surveillance, adaptive compressing, detecting unknown objects and scene understanding. In this paper, we propose a small and sparse neural network model for performing salient object segmentation that is suitable for use in mobile and embedded applications. Our model is built using depthwise separable convolutions and bottleneck inverted residuals which have been proven to perform very memory-efficient inference and can be easily implemented using standard functions available in all deep learning frameworks. The multiscale features extracted along with the layers with deep residuals allow our network to learn high-quality saliency maps. We present the quantitative results of our QUICKSAL model with multiple levels of model sparsity ranging from 0% to 96%, with the non-zero parameter count varying from 3.3M to 0.14M respectively - on publicly available benchmark datasets - showing that our highly constrained approach is comparable to other state-of-the-art approaches (parameter count 35M). We also present qualitative results on camouflage images and show that our model can successfully distinguish between the salient and non-salient parts even when both seem blended together.

****

MonoLayout: Amodal scene layout from a single image

Kaustubh Mani, Swapnil Daga, Shubhika Garg, Sai Shankar Narasimhan, Madhava Krishna, Krishna Murthy Jatavallabhula; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1689-1697

In this paper, we address the novel, highly challenging problem of estimating the layout of a complex urban driving scenario. Given a single color image captured from a driving platform, we aim to predict the bird's eye view layout of the road and other traffic participants. The estimated layout should reason beyond what is visible in the image, and compensate for the loss of 3D information due to projection. We dub this problem amodal scene layout estimation, which involves hallucinating scene layout for even parts of the world that are occluded in the image. To this end, we present MonoLayout, a deep neural network for real- time amodal scene layout estimation from a single image. We represent scene layout as a multi-channel semantic occupancy grid, and leverage adversarial feature learning to "hallucinate" plausible completions for occluded image parts. We extend several state-of-the-art approaches for road-layout estimation and vehicle occupancy estimation in bird's eye view to the amodal setup and thoroughly evaluate against them. By leveraging temporal sensor fusion to generate training labels, we significantly outperform current art (> 10% improvement) over a number of datasets. We also make all our annotations, code, and pretrained models publicly available.

****

Frustum VoxNet for 3D object detection from RGB-D or Depth images

Xiaoke Shen, Ioannis Stamos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1698-1706

Recently, there have been a plethora of classification and detection systems from RGB as well as 3D images. In this work, we describe a new 3D object detection system from an RGB-D or depth-only point cloud. Our system first detects objects in 2D (either RGB, or pseudo-RGB constructed from depth). The next step is to detect 3D objects within the 3D frustums these 2D detections define. This is achieved by voxelizing parts of the frustums (since frustums can be really large), instead of using the whole frustums as done in earlier work. The main novelty of our system has to do with determining which parts (3D proposals) of the frustums to voxelize, thus allowing us to provide high resolution representations around the objects of interest. It also allows our system to have reduced memory requirements. These 3D proposals are fed to an efficient ResNet-based 3D Fully Convolutional Network (FCN). Our 3D detection system is fast, and can be integrated into a robotics platform. With respect to systems that do not perform voxelization (such as PointNet), our methods can operate without the requirement of subsampling of the datasets. We have also introduced a pipelining approach that further improves the efficiency of our system. Results on SUN RGB-D dataset show that our system, which is based on a small network, can process 20 frames per second with comparable detection results to the state-of-the-art [16], achieving a 2x spe edup.

****

Cross-View Contextual Relation Transferred Network for Unsupervised Vehicle Tracking in Drone Videos

Wenfeng Song, Shuai Li, Tao Chang, Aimin Hao, Qinping Zhao, Hong Qin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1707-1716

Recently CNN-centric object tracking methods have been gaining tremendous success in ground-view videos, however, it remains hard to cope with vehicle tracking in unmanned aerial vehicle (UAV) videos. The key difficulties mainly stem from lacking large-scale well-labeled training datasets and view-invariant appearance model for fast-moving drone-view vehicles. We enhance the vehicle's cross-view feature by exploring relations between the pivotal context and the target to facilitate unsupervised vehicle tracking. The relation is modeled as the relevance of the target and its contextual regions in the tracking task.   Specifically, we propose a contextual relation actor-critic (CRAC) framework integrates an actor-critic agent with a dual GAN learning mechanism, which aims to dynamically search the related contextual regions and transfer the relations from ground-view to drone-view videos while retaining the discriminative features.   We demonstrate that CRAC could be applied to several state-of-the-art trackers by extensive experiments and ablation studies on four public benchmarks. All the experiments confirm that, our CRAC can improve the performance of state-of-the-art methods in terms of accuracy, robustness, and versatility.

****

Unsupervised and Semi-Supervised Domain Adaptation for Action Recognition from Drones

Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, Jia-Bin Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1717-1726

We address the problem of human action classification in drone videos. Due to the high cost of capturing and labeling large-scale drone videos with diverse actions, we present unsupervised and semi-supervised domain adaptation approaches that leverage both the existing fully annotated action recognition datasets and unannotated (or only a few annotated) videos from drones. To study the emerging problem of drone-based action recognition, we create a new dataset, NEC-Drone, containing 5,250 videos to evaluate the task. We tackle both problem settings with 1) same and 2) different action label sets for the source (e.g., Kinectics dataset) and target domains (drone videos). We present a combination of video and instance-based adaptation methods, paired with either a classifier or an embedding-based framework to transfer the knowledge from source to target. Our results show that the proposed adaptation approach substantially improves the performance on these challenging and practical tasks. We further demonstrate the applicability of our method for learning cross-view action recognition on the Charades-Ego dataset. We provide qualitative analysis to understand the behaviors of our approaches.

****

Localizing Grouped Instances for Efficient Detection in Low-Resource Scenarios

Amelie Royer, Christoph Lampert; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1727-1736

State-of-the-art detection systems generally focus, and are evaluated on, their ability to exhaustively retrieve objects densely distributed in the image, across a wide variety of appearances and semantic categories. Orthogonal to this, many practical object detection applications, for example in remote sensing, instead require dealing with large images that contain only a few small objects of a single class, scattered heterogeneously across the space. In addition, they are often subject to strict computational constraints, such as limited battery capacity and computing power. To tackle these more practical scenarios, we propose a novel detection scheme that offers a flexible and efficient framework for detection tasks with variable object sizes and densities: We rely on a sequence of detection stages, each of which has the ability to predict groups of objects as wel

l as individuals. Similar to a detection cascade, this multi-stage architecture spares computational effort by discarding large irrelevant regions of the image early during the detection process. The ability to group objects provides further computational and memory savings, as it allows working with lower image resolutions in early stages, where groups are more easily detected than individuals. We report experimental results on two aerial image datasets, and show that the proposed method is as accurate yet computationally more efficient than standard single-shot detectors, consistently across three different backbone architectures.
****

Reconstructing Road Network Graphs from both Aerial Lidar and Images
Biswas Parajuli, Ahana Roy Choudhury, Piyush Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1737-1746
We address the problem of reconstructing road networks as undirected graphs over large geographic regions in cold start scenarios where neither the preliminary graph nor any on-road trajectory information is available. The goal of this paper is to transform bimodal aerial data in the form of 3-dimensional Lidar scans and high resolution images into road network graphs. We use a fully convolutional architecture that fuses the two datasets by reducing the disparity in their modalities to segment out roads. We then apply a simple, disk-packing based algorithm that covers the segmented regions with a minimal set of variably sized disks, connect the intersecting disks and use a provable curve reconstruction algorithm to obtain the road network graph. We show that our method is better at removing outliers and gives improved connectivity and topological accuracy than the existing state of the art thinning based method.
****

BIRDSAI: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos
Elizabeth Bondi, Raghav Jain, Palash Aggrawal, Saket Anand, Robert Hannaford, Ashish Kapoor, Jim Piavis, Shital Shah, Lucas Joppa, Bistra Dilkina, Milind Tambe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1747-1756
Monitoring of protected areas to curb illegal activities like poaching and animal trafficking is a monumental task. To augment existing manual patrolling efforts, unmanned aerial surveillance using visible and thermal infrared (TIR) cameras is increasingly being adopted. Automated data acquisition has become easier with advances in unmanned aerial vehicles (UAVs) and sensors like TIR cameras, which allow surveillance at night when poaching typically occurs. However, it is still a challenge to accurately and quickly process large amounts of the resulting TIR data. In this paper, we present the first large dataset collected using a TIR camera mounted on a fixed-wing UAV in multiple African protected areas. This dataset includes TIR videos of humans and animals with several challenging scenarios like scale variations, background clutter due to thermal reflections, large camera rotations, and motion blur. Additionally, we provide another dataset with videos synthetically generated with the publicly available Microsoft AirSim simulation platform using a 3D model of an African savanna and a TIR camera model. Through our benchmarking experiments on state-of-the-art detectors, we demonstrate that leveraging the synthetic data in a domain adaptive setting can significantly improve detection performance. We also evaluate various recent approaches for single and multi-object tracking. With the increasing popularity of aerial imagery for monitoring and surveillance purposes, we anticipate this unique dataset to be used to develop and evaluate techniques for object detection, tracking, and domain adaptation for aerial, TIR videos.
****

Dual-Mode Training with Style Control and Quality Enhancement for Road Image Domain Adaptation
Moritz Venator, Fengyi Shen, Selcuk Aklanoglu, Erich Bruns, Klaus Diepold, Andreas Maier; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1757-1766
Dealing properly with different viewing conditions remains a key challenge for computer vision in autonomous driving. Domain adaptation has opened new possibili

ties for data augmentation, translating arbitrary road scene images into different environmental conditions. Although multimodal concepts have demonstrated the capability to separate content and style, we find that existing methods fail to reproduce scenes in the exact appearance given by a reference image. In this paper, we address the aforementioned problem by introducing a style alignment loss between output and reference image. We integrate this concept into a multimodal unsupervised image-to-image translation model with a novel dual-mode training process and additional adversarial losses. Focusing on road scene images, we evaluate our model in various aspects including visual quality and feature matching. Our experiments reveal that we are able to significantly improve both style alignment and image quality in different viewing conditions. Adapting concepts from neural style transfer, our new training approach allows to control the output of multimodal domain adaptation, making it possible to generate arbitrary scenes and viewing conditions for data augmentation.
****

Periphery-Fovea Multi-Resolution Driving Model Guided by Human Attention
Ye Xia, Jinkyu Kim, John Canny, Karl Zipser, Teresa Canas-Bajo, David Whitney; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1767-1775
Inspired by human vision, we propose a new periphery-fovea multi-resolution driving model that predicts vehicle speed from dash camera videos. The peripheral vision module of the model processes the full video frames in low resolution with large receptive fields. Its foveal vision module selects sub-regions and uses high-resolution input from those regions to improve its driving performance. We train the fovea selection module with supervision from driver gaze. We show that adding high-resolution input from predicted human driver gaze locations significantly improves the driving accuracy of the model. Our periphery-fovea multi-resolution model outperforms a uni-resolution periphery-only model that has the same amount of floating-point operations. More importantly, we demonstrate that our driving model achieves a significantly higher performance gain in pedestrian-involved critical situations than in other non-critical situations. Our code is publicly available at https://github.com/pascalxia/periphery_fovea_driving.
****

Deep Remote Sensing Methods for Methane Detection in Overhead Hyperspectral Imagery
Satish Kumar, Carlos Torres, Oytun Ulutan, Alana Ayasse, Dar Roberts, B. S. Manjunath; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1776-1785
Effective analysis of hyperspectral imagery is essential for gathering fast and actionable information of large areas affected by atmospheric and green house gases. Existing methods, which process hyperspectral data to detect amorphous gases such as CH4 require manual inspection from domain experts and annotation of massive datasets. These methods do not scale well and are prone to human errors due to the plumes' small pixel-footprint signature. The proposed Hyperspectral Mask-RCNN (H-mrcnn) uses principled statistics, signal processing, and deep neural networks to address these limitations. H-mrcnn introduces fast algorithms to analyze large-area hyper-spectral information and methods to autonomously represent and detect CH4 plumes. H-mrcnn processes information by match-filtering sliding windows of hyperspectral data across the spectral bands. This process produces information-rich features that are both effective plume representations and gas concentration analogs. The optimized matched-filtering stage processes spectral data, which is spatially sampled to train an ensemble of gas detectors. The ensemble outputs are fused to estimate a natural and accurate plume mask. Thorough evaluation demonstrates that H-mrcnn matches the manual and experience-dependent annotation process of experts by 85% (IOU). H-mrcnn scales to larger datasets, reduces the manual data processing and labeling time (12 times), and produces rapid actionable information about gas plumes.
****

City-Scale Road Extraction from Satellite Imagery v2: Road Speeds and Travel Times

Adam Van Etten; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1786-1795

Automated road network extraction from remote sensing imagery remains a significant challenge despite its importance in a broad array of applications. To this end, we explore road network extraction at scale with inference of semantic features of the graph, identifying speed limits and route travel times for each roadway. We call this approach City-Scale Road Extraction from Satellite Imagery v2 (CRESIv2), Including estimates for travel time permits true optimal routing (rather than just the shortest geographic distance), which is not possible with existing remote sensing imagery based methods. We evaluate our method using two sources of labels (OpenStreetMap, and those from the SpaceNet dataset), and find that models both trained and tested on SpaceNet labels outperform OpenStreetMap labels by greater than 60%. We quantify the performance of our algorithm with the Average Path Length Similarity (APLS) and map topology (TOPO) graph-theoretic metrics over a diverse test area covering four cities in the SpaceNet dataset. For a traditional edge weight of geometric distance, we find an aggregate of 5% improvement over existing methods for SpaceNet data. We also test our algorithm on Google satellite imagery with OpenStreetMap labels, and find a 23% improvement over previous work. Metric scores decrease by only 4% on large graphs when using travel time rather than geometric distance for edge weights, indicating that optimizing routing for travel time is feasible with this approach.

****

Cloud Removal from Satellite Images using Spatiotemporal Generator Networks

Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, Stefano Ermon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1796-1805

Satellite images hold great promise for continuous environmental monitoring and earth observation. Occlusions cast by clouds, however, can severely limit coverage, making ground information extraction more difficult. Existing pipelines typically perform cloud removal with simple temporal composites and hand-crafted filters. In contrast, we cast the problem of cloud removal as a conditional image synthesis challenge, and we propose a trainable spatiotemporal generator network (STGAN) to remove clouds. We train our model on a new large-scale spatiotemporal dataset that we construct, containing 97640 image pairs covering all continents. We demonstrate experimentally that the proposed STGAN model outperforms standard models and can generate realistic cloud-free images with high PSNR and SSIM values across a variety of atmospheric conditions, leading to improved performance in downstream tasks such as land cover classification.

****

Single Satellite Optical Imagery Dehazing using SAR Image Prior Based on conditional Generative Adversarial Networks

Binghui Huang, Li Zhi, Chao Yang, Fuchun Sun, Yixu Song; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1806-1813

Satellite image dehazing aims at precisely retrieving the real situations of the obscured parts from the hazy remote sensing (RS) images, which is a challenging task since the hazy regions contain both ground features and haze components. Many approaches of removing haze focus on processing multi-spectral or RGB images, whereas few of them utilize multi-sensor data. The multi-sensor data fusion is significant to provide auxiliary information since RGB images are sensitive to atmospheric conditions. In this paper, a dataset called SateHaze1k is established and composed of 1200 pairs clear Synthetic Aperture Radar (SAR), hazy RGB, and corresponding ground truth images, which are divided into three degrees of the haze, i.e. thin, moderate, and thick fog. Moreover, we propose a novel fusion dehazing method to directly restore the haze-free RS images by using an end-to-end conditional generative adversarial network(cGAN). The proposed network combines the information of both RGB and SAR images to eliminate the image blurring. Besides, the dilated residual blocks of the generator can also sufficiently improve the dehazing effects. Our experiments demonstrate that the proposed method, which fuses the information of different sensors applied to the cloudy conditions, can achieve more precise results than other baseline models.

****

The Synthinel-1 dataset: a collection of high resolution synthetic overhead imagery for building segmentation

Fanjie Kong, Bohao Huang, Kyle Bradbury, Jordan Malof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1814-1823

Recently deep learning - namely convolutional neural networks (CNNs) - have yielded impressive performance for the task of building segmentation on large overhead (e.g., satellite) imagery benchmarks. However, these benchmark datasets only capture a small fraction of the variability present in real-world overhead imagery, limiting the ability to properly train, or evaluate, models for real-world application. Unfortunately, developing a dataset that captures even a small fraction of real-world variability is typically infeasible due to the cost of imagery, and manual pixel-wise labeling of the imagery. In this work we develop an approach to rapidly and cheaply generate large and diverse synthetic overhead imagery for training segmentation CNNs. Using this approach, we generate and publicly-release a collection of synthetic overhead imagery, termed Synthinel-1, with full pixel-wise building labels. We use several benchmark datasets to demonstrate that Synthinel-1 is consistently beneficial when used to augment real-world training imagery, especially when CNNs are tested on novel geographic locations or conditions.

****

Efficient Object Detection in Large Images Using Deep Reinforcement Learning

Burak Uzkent, Christopher Yeh, Stefano Ermon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1824-1833

Traditionally, an object detector is applied to every part of the scene of interest, and its accuracy and computational cost increases with higher resolution images. However, in some application domains such as remote sensing, purchasing high spatial resolution images is expensive. To reduce the large computational and monetary cost associated with using high spatial resolution images, we propose a conditional reinforcement learning agent that adaptively selects the spatial resolution of each image that is provided to the detector. In particular, we train the agent in a dual reward setting to choose low spatial resolution images to be run through a coarse level detector when the image is dominated by large objects, and high spatial resolution image to be run through a fine level detector when it is dominated by small objects. This reduces the dependency on high spatial resolution images for building a robust detector and increases run-time efficiency. We perform experiments on the xView dataset, consisting of large images, where we increase run-time efficiency by 60% and use high resolution images only 30% of the time while maintaining similar accuracy as a detector that uses only high resolution images.

****

Lane detection using lane boundary marker network with road geometry constraints

Hussam Ullah Khan, Afsheen Rafaqat Ali, Ali Hassan, Ahmed Ali, Wajahat Kazmi, Aamer Zaheer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1834-1843

Lane detection is of critical importance to both the self-driving cars as well as advanced driver assistance systems. While current methods use a range of features from low-level to deep features extracted from convolutional neural networks, they all suffer from the problem of occlusion and struggle to detect lanes with low or no evidence on the road. In this paper, we use a lane boundary marker network to detect keypoints along the lane boundaries. An inverse perspective mapping is estimated using road geometry which is then applied to the detected markers and lines/curves are fitted jointly on the rectified points. Finally, missing lane boundaries are predicted using lane geometry constraints i.e., equidistant and parallelism. Reciprocal weighted averaging ensures lane boundaries with strong evidence dominate their predicted alternatives. The results show a significant improvement of +7.8%, +6.8% and +1.2% of F1 scores over the state-of-the-art on CULane, Caltech and TuSimple datasets, respectively. This proves our algorit hm's robustness against both occluded and missing lanes cases. Furthermore, we also show that our algorithm can be combined with other lane detectors to improve their lane retrieval potential.

****

It's All About The Scale - Efficient Text Detection Using Adaptive Scaling

Elad Richardson, Yaniv Azar, Or Avioz, Niv Geron, Tomer Ronen, Zach Avraham, Stav Shapiro; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1844-1853

"Text can appear anywhere". This property requires us to carefully process all the pixels in an image in order to accurately localize all text instances. In particular, for the more difficult task of localizing small text regions, many methods use an enlarged image or even several rescaled ones as their input. This significantly increases the processing time of the entire image and needlessly enlarges background regions. If we were to have a prior telling us the coarse location of text instances in the image and their approximate scale, we could have adaptively chosen which regions to process and how to rescale them, thus significantly reducing the processing time. To estimate this prior we propose a segmentation-based network with an additional "scale predictor", an output channel that predicts the scale of each text segment. The network is applied on a scaled down image to efficiently approximate the desired prior, without processing all the pixels of the original image. The approximated prior is then used to create a compact image containing only text regions, resized to a canonical scale, which is fed again to the segmentation network for fine-grained detection. We show that our approach offers a powerful alternative to fixed scaling schemes, achieving an equivalent accuracy to larger input scales while processing far fewer pixels. Qualitative and quantitative results are presented on the ICDAR15 and ICDAR17 MLT benchmarks to validate our approach.

****

Casting Geometric Constraints in Semantic Segmentation as Semi-Supervised Learning

Sinisa Stekovic, Friedrich Fraundorfer, Vincent Lepetit; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1854-1863

We propose a simple yet effective method to learn to segment new indoor scenes from video frames: State-of-the-art methods trained on one dataset, even as large as the SUNRGB-D dataset, can perform poorly when applied to images that are not part of the dataset, because of the dataset bias, a common phenomenon in computer vision. To make semantic segmentation more useful in practice, one can exploit geometric constraints. Our main contribution is to show that these constraints can be cast conveniently as semi-supervised terms, which enforce the fact that the same class should be predicted for the projections of the same 3D location in different images. This is interesting as we can exploit general existing techniques developed for semi-supervised learning to efficiently incorporate the constraints. We show that this approach can efficiently and accurately learn to segment target sequences of ScanNet and our own target sequences using only annotations from SUNRGB-D, and geometric relations between the video frames of target sequences.

****

MLSL: Multi-Level Self-Supervised Learning for Domain Adaptation with Spatially Independent and Semantically Consistent Labeling

Javed Iqbal, Mohsen Ali; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1864-1873

Most of the recent Deep Semantic Segmentation algorithms suffer from large generalization errors, even when powerful hierarchical representation models, based on convolutional neural networks, have been employed. This could be attributed to limited training data and large distribution gap in train and test domain datasets. In this paper, we propose a multi-level self-supervised learning model for domain adaptation of semantic segmentation. Exploiting the idea that an object (and most of stuff given context) should be labeled consistently regardless of its location, we generate spatially independent and semantically consistent (SISC)

pseudo-labels by segmenting multiple sub-images using base model and designing an aggregation strategy. Image level pseudo weak-labels, PWL, are computed to guide domain adaptation by capturing global context similarity in source and target domain at latent space level. Thus helping latent space learn the representation even when there are very few pixels belonging to the domain category (small object for example) compared to rest of the image. Our multi-level Self-supervised learning (MLSL) outperforms existing state-of-art (self or adversarial learning) algorithms. Specifically, keeping all setting similar and employing MLSL we obtain a mIoU gain of 5.1% on GTA-V to Cityscapes adaptation and 4.3% on SYNTHIA to Cityscapes adaptation compared to the existing state-of-art method
****

FuseSeg: LiDAR Point Cloud Segmentation Fusing Multi-Modal Data
Georg Krispel, Michael Opitz, Georg Waltner, Horst Possegger, Horst Bischof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1874-1883
We introduce a simple yet effective fusion method of LiDAR and RGB data to segment LiDAR point clouds. Utilizing the dense native range representation of a LiDAR sensor and the setup calibration, we establish point correspondences between the two input modalities. Subsequently, we are able to warp and fuse the features from one domain into the other. Therefore, we can jointly exploit information from both data sources within one single network. To show the merit of our method, we extend SqueezeSeg, a point cloud segmentation network, with an RGB feature branch and fuse it into the original structure. Our extension called FuseSeg leads to an improvement of up to 18% IoU on the KITTI benchmark. In addition to the improved accuracy, we also achieve real-time performance at 50 fps, five times as fast as the KITTI LiDAR data recording speed.
****

EpO-Net: Exploiting Geometric Constraints on Dense Trajectories for Motion Saliency
Ijaz Akhter, Mohsen Ali, Muhammad Faisal, RICHARD HARTLEY; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1884-1893
The existing approaches for salient motion segmentation are unable to explicitly learn geometric cues and often give false detections on prominent static objects. We exploit multiview geometric constraints to avoid such shortcomings. To handle the nonrigid background like a sea, we also propose a robust fusion mechanism between motion and appearance-based features. We find dense trajectories, covering every pixel in the video, and propose trajectory-based epipolar distances to distinguish between background and foreground regions. Trajectory epipolar distances are data-independent and can be readily computed given a few features' correspondences between the images. We show that by combining epipolar distances with optical flow, a powerful motion network can be learned. Enabling the network to leverage both of these features, we propose a simple mechanism, we call input-dropout. Comparing the motion-only networks, we outperform the previous state of the art on DAVIS-2016 dataset by 5.2% in the mean IoU score. By robustly fusing our motion network with an appearance network using the input-dropout mechanism, we also outperform the previous methods on DAVIS-2016, 2017 and Segtrackv2 dataset.
****

Multi Receptive Field Network for Semantic Segmentation
Jianlong Yuan, Zelu Deng, Shu Wang, Zhenbo Luo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1894-1903
Semantic segmentation is one of the key tasks in computer vision, which is to assign a category label to each pixel in an image. Despite significant progress achieved recently, most existing methods still suffer from two challenging issues: 1) the size of objects and stuff in an image can be very diverse, demanding for incorporating multi-scale features into the fully convolutional networks (FCNs); 2) the pixels close to or at the boundaries of object/stuff are hard to classify due to the intrinsic weakness of convolutional networks. To address the first issue, we propose a new Multi-Receptive Field Module (MRFM), explicitly taking

multi-scale features into account. For the second issue, we design an edge-aware loss which is effective in distinguishing the boundaries of object/stuff. With these two designs, our Multi Receptive Field Network achieves new state-of-the-art results on two widely-used semantic segmentation benchmark datasets. Specifically, we achieve a mean IoU of 83.0% on the Cityscapes dataset and 88.4% mean IoU on the pascal VOC2012 dataset.

****

DIPNet: Dynamic Identity Propagation Network for Video Object Segmentation

Ping Hu, Jun Liu, Gang Wang, Vitaly Ablavsky, Kate Saenko, Stan Sclaroff; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1904-1913

Many recent methods for semi-supervised Video Object Segmentation (VOS) have achieved good performance by exploiting the annotated first frame via one-shot fine-tuning or mask propagation. However, heavily relying on the first frame may weaken the robustness for VOS, since video objects can show large variations through time. In this work, we propose a Dynamic Identity Propagation Network (DIPNet) that adaptively propagates and accurately segments the video objects over time. To achieve this, DIPNet disentangles the VOS task at each time step into a dynamic propagation phase and a spatial segmentation phase. The former utilizes a novel identity representation to adaptively propagate objects' reference information over time, which enhances the robustness to video objects' temporal variations. The latter uses the propagated information to tackle the object segmentation as an easier static image problem that can be optimized via slight fine-tuning on the first frame, thus reducing the computational cost. As a result, by optimizing these two components to complement each other, we can achieve a robust system for VOS. Evaluations on four benchmark datasets show that DIPNet provides state-of-the-art performance with time efficiency.

****

Representing Objects in Video as Space-Time Volumes by Combining Top-Down and Bottom-Up Processes

Filip Ilic, Axel Pinz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1914-1922

As top-down based approaches of object recognition from video are getting more powerful, a structured way to combine them with bottom-up grouping processes becomes feasible. When done right, the resulting representation is able to describe objects and their decomposition into parts at appropriate spatio-temporal scales.We propose a method that uses a modern object detector to focus on salient structures in video, and a dense optical flow estimator to supplement feature extraction. From these structures we extract space-time volumes of interest (STVIs) by smoothing in spatio-temporal Gaussian Scale Space that guides bottom-up grouping.The resulting novel representation enables us to analyze and visualize the decomposition of an object into meaningful parts while preserving temporal object continuity. Our experimental validation is twofold. First, we achieve competitive results on a common video object segmentation benchmark. Second, we extend this benchmark with high quality object part annotations, DAVIS Parts, on which we establish a strong baseline by showing that our method yields spatio-temporally meaningful object parts. Our new representation will support applications that require high-level space-time reasoning at the parts level.

****

Dense Extreme Inception Network: Towards a Robust CNN Model for Edge Detection

Xavier Soria Poma, Edgar Riba, Angel Sappa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1923-1932

This paper proposes a Deep Learning based edge detector, which is inspired on both HED (Holistically-Nested Edge Detection) and Xception networks. The proposed approach generates thin edge-maps that are plausible for human eyes; it can be used in any edge detection task without previous training or fine tuning process. As a second contribution, a large dataset with carefully annotated edges, has been generated. This dataset has been used for training the proposed approach as well the state-of-the-art algorithms for comparisons. Quantitative and qualitative evaluations have been performed on different benchmarks showing improvements with the proposed method when F-measure of ODS and OIS are considered.

****

Can I teach a robot to replicate a line art
Raghav B.V., Subham Kumar, Vinay Namboodiri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1933-1941

Line art is arguably one of the fundamental and versatile modes of expression. We propose a pipeline for a robot to look at a grayscale line art and redraw it. The key novel elements of our pipeline are: a) we propose a novel task of mimicking line drawings, b) to solve the pipeline we modify the Quick-draw dataset and obtain supervised training for converting a line drawing into a series of strokes c) we propose a multi-stage segmentation and graph interpretation pipeline for solving the problem. The resultant method has also been deployed on a CNC plotter as well as a robotic arm. We have trained several variations of the proposed methods and evaluate these on a dataset obtained from Quick-draw. Through the best methods we observe an accuracy of around 98% for this task, which is a significant improvement over the baseline architecture we adapted from. This therefore allows for deployment of the method on robots for replicating line art in a reliable manner. We also show that while the rule-based vectorization methods do suffice for simple drawings, it fails for more complicated sketches, unlike our method which generalizes well to more complicated distributions.

****

Multiview Co-segmentation for Wide Baseline Images using Cross-view Supervision
Yuan Yao, Hyun Soo Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1942-1951

This paper presents a method to co-segment an object from wide baseline multiview images using cross-view self-supervision. A key challenge in the wide baseline images lies in the fragility of photometric matching. Inspired by shape-from-silhouette that does not require photometric matching, we formulate a new theory of shape belief transfer---the segmentation belief in one image can be used to predict that of the other image through epipolar geometry. This formulation is differentiable, and therefore, an end-to-end training is possible. We analyze the shape belief transfer to identify the theoretical upper and lower bounds of the unlabeled data segmentation, which characterizes the degenerate cases of co-segmentation. We design a novel triple network that embeds this shape belief transfer, which is agnostic to visual appearance and baseline. The resulting network is validated by recognizing a target object from realworld visual data including non-human species and a subject of interest in social videos where attaining large-scale annotated data is challenging.

****

Shape Constrained Network for Eye Segmentation in the Wild
Bingnan Luo, Jie Shen, Shiyang Cheng, Yujiang Wang, Maja Pantic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1952-1960

Semantic segmentation of eyes has long been a vital pre-processing step in many biometric applications. Majority of the works focus only on high resolution eye images, while little has been done to segment the eyes from low quality images in the wild. However, this is a particularly interesting and meaningful topic, as eyes play a crucial role in conveying the emotional state and mental well-being of a person. In this work, we take two steps toward solving this problem: (1) We collect and annotate a challenging eye segmentation dataset containing 8882 eye patches from 4461 facial images of different resolutions, illumination conditions and head poses; (2) We develop a novel eye segmentation method, Shape Constrained Network (SCN), that incorporates shape prior into the segmentation network training procedure. Specifically, we learn the shape prior from our dataset using VAE-GAN, and leverage the pre-trained encoder and discriminator to regularise the training of SegNet. To improve the accuracy and quality of predicted masks, we replace the loss of SegNet with three new losses: Intersection-over-Union (IoU) loss, shape discriminator loss and shape embedding loss. Extensive experiments shows that our method outperforms state-of-the-art segmentation and landmark detection methods in terms of mean IoU (mIoU) accuracy and the quality of segmen tation masks. The dataset is available at https://ibug.doc.ic.ac.uk/resources/ibug-eye-segmentation-dataset/

****

ROSS: Robust Learning of One-shot 3D Shape Segmentation

Shuaihang Yuan, Yi Fang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1961-1969

3D shape segmentation is a fundamental computer vision task that partitions the object into labeled semantic parts. Recent approaches to 3D shape segmentation learning heavily rely on high-quality labeled training datasets. This limits their use in applications to handle the large scale unannotated datasets. In this paper, we proposed a novel semi-supervised approach, named Robust Learning of One-Shot 3D Shape Segmentation (ROSS), which only requires one single exemplar labeled shape for training. The proposed ROSS can generalize its ability from a one-shot training process to predict the segmentation for previously unseen 3D shape models. The proposed ROSS is composed of three major modules for 3D shape segmentation as follows. The global shape descriptor generator is the first module that utilizes the proposed reference weighted convolution to learn a 3D shape descriptor. The second module is a part-aware shape descriptor constructor that can generate weighted descriptors from a learned 3D shape descriptor according to semantic parts without supervision. The shape morphing with label transferring works as the last module. It morphs the exemplar shape and then transfers labels from the transformed exemplar shape to the target shape. The extensive experimental results on 3D mesh datasets demonstrate the ROSS is robust to noise and incomplete shapes and it can be applied to unannotated datasets. The experiment shows the proposed ROSS can achieve comparable performance with the supervised method.

****

Architecture Search of Dynamic Cells for Semantic Video Segmentation

Vladimir Nekrasov, Hao Chen, Chunhua Shen, Ian Reid; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1970-1979

In semantic video segmentation the goal is to acquire consistent dense semantic labelling across image frames. To this end, recent approaches have been reliant on manually arranged operations applied on top of static semantic segmentation networks -- with the most prominent building block being the optical flow able to provide information about scene dynamics. Related to that is the line of research concerned with speeding up static networks by approximating expensive parts of them with cheaper alternatives, while propagating information from previous frames. In this work we attempt to come up with generalisation of those methods, and instead of manually designing contextual blocks that connect per-frame outputs, we propose a neural architecture search solution, where the choice of operations together with their sequential arrangement are being predicted by a separate neural network. We showcase that such generalisation leads to stable and accurate results across common benchmarks, such as CityScapes and CamVid datasets. Importantly, the proposed methodology takes only 2 GPU-days, finds high-performing cells and does not rely on the expensive optical flow computation.

****

Template-Based Automatic Search of Compact Semantic Segmentation Architectures

Vladimir Nekrasov, Chunhua Shen, Ian Reid; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1980-1989

Automatic search of neural architectures for various vision and natural language tasks is becoming a prominent tool as it allows to discover high-performing structures on any dataset of interest. Nevertheless, on more difficult domains, such as dense per-pixel classification, current automatic approaches are limited in their scope -- due to their strong reliance on existing image classifiers they tend to search only for a handful of additional layers with discovered architectures still containing a large number of parameters. In contrast, in this work we propose a novel solution able to find light-weight and accurate segmentation architectures starting from only few blocks of a pre-trained classification network. To this end, we progressively build up a methodology that relies on templates of sets of operations, predicts which template and how many times should be app

lied at each step, while also generating the connectivity structure and downsampling factors. All these decisions are being made by a recurrent neural network that is rewarded based on the score of the emitted architecture on the holdout set and trained using reinforcement learning. One discovered architecture achieves 63.2% mean IoU on CamVid and 67.8% on CityScapes having only 270K parameters.
****

Simultaneous Detection and Removal of Dynamic Objects in Multi-view Images
Gagan Kanojia, Shanmuganathan Raman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 1990-1999
Consider a set of images of a scene consisting of moving objects captured using a hand-held camera. In this work, we propose an algorithm which takes this set of multi-view images as input, detects the dynamic objects present in the scene, and replaces them with the static regions which are being occluded by them. The proposed algorithm scans the reference image in the row-major order at the pixel level and classifies each pixel as static or dynamic. During the scan, when a pixel is classified as dynamic, the proposed algorithm replaces that pixel value with the corresponding pixel value of the static region which is being occluded by that dynamic region. We show that we achieve artifact-free removal of dynamic objects in multi-view images of several real-world scenes. To the best of our knowledge, we propose the first method which simultaneously detects and removes the dynamic objects present in multi-view images.
****

UnOVOST: Unsupervised Offline Video Object Segmentation and Tracking
Jonathon Luiten, Idil Esen Zulfikar, Bastian Leibe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2000-2009
We address Unsupervised Video Object Segmentation (UVOS), the task of automatically generating accurate pixel masks for salient objects in a video sequence and of tracking these objects consistently through time, without any input about which objects should be tracked. Towards solving this task, we present UnOVOST (Unsupervised Offline Video Object Segmentation and Tracking) as a simple and generic algorithm which is able to track and segment a large variety of objects. This algorithm builds up tracks in a number stages, first grouping segments into short tracklets that are spatio-temporally consistent, before merging these tracklets into long-term consistent object tracks based on their visual similarity. In order to achieve this we introduce a novel tracklet-based Forest Path Cutting data association algorithm which builds up a decision forest of track hypotheses before cutting this forest into paths that form long-term consistent object tracks. When evaluating our approach on the DAVIS 2017 Unsupervised dataset we obtain state-of-the-art performance with a mean J &F score of 67.9% on the val, 58% on the test-dev and 56.4% on the test-challenge benchmarks, obtaining first place in the DAVIS 2019 Unsupervised Video Object Segmentation Challenge. UnOVOST even performs competitively with many semi-supervised video object segmentation algorithms even though it is not given any input as to which objects should be tracked and segmented.
****

Leveraging Pretrained Image Classifiers for Language-Based Segmentation
David Golub, Roberto Martin-Martin, Ahmed El-Kishky, Silvio Savarese; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2010-2019
Current semantic segmentation models cannot easily generalize to new object classes unseen during train time: they require additional annotated images and retraining. We propose a novel segmentation model that injects visual priors from pretrained image classifiers into semantic segmentation architectures, allowing them to segment out new target labels without retraining. As visual priors, we use the activations of pretrained image classifiers, which provide noisy indications of the spatial location of both the target object and distractor objects in the scene. We leverage language semantics to obtain these activations for a target label unseen by the classifier. Further experiments show that the visual priors obtained via language semantics for both relevant anddistracting objects are key

to our performance

****

Quadtree Generating Networks: Efficient Hierarchical Scene Parsing with Sparse Convolutions

Kashyap Chitta, Jose M. Alvarez, Martial Hebert; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2020-2029

Semantic segmentation with Convolutional Neural Networks is a memory-intensive task due to the high spatial resolution of feature maps and output predictions. In this paper, we present Quadtree Generating Networks (QGNs), a novel approach able to drastically reduce the memory footprint of modern semantic segmentation networks. The key idea is to use quadtrees to represent the predictions and target segmentation masks instead of dense pixel grids. Our quadtree representation enables hierarchical processing of an input image, with the most computationally demanding layers only being used at regions in the image containing boundaries between classes. In addition, given a trained model, our representation enables flexible inference schemes to trade-off accuracy and computational cost, allowing the network to adapt in constrained situations such as embedded devices. We demonstrate the benefits of our approach on the Cityscapes, SUN-RGBD and ADE20k datasets. On Cityscapes, we obtain an relative 3% mIoU improvement compared to a dilated network with similar memory consumption; and only receive a 3% relative mIoU drop compared to a large dilated network, while reducing memory consumption by over 4x. Our code is available at https://github.com/kashyap7x/QGN.

****

MaskPlus: Improving Mask Generation for Instance Segmentation

Shichao Xu, Shuyue Lan, Zhu Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2030-2038

Instance segmentation is a promising yet challenging topic in computer vision. Recent approaches such as Mask R-CNN typically divide this problem into two parts -- a detection component and a mask generation branch, and mostly focus on the improvement of the detection part.   In this paper, we present an approach that extends Mask R-CNN with five novel techniques for improving the mask generation branch and reducing the conflicts between the mask branch and the detection component in training.   These five techniques are independent to each other and can be flexibly utilized in building various instance segmentation architectures for increasing the overall accuracy. We demonstrate the effectiveness of our approach with tests on the COCO dataset.

****

Multi-Level Representation Learning for Deep Subspace Clustering

Mohsen Kheirandishfard, Fariba Zohrizadeh, Farhad Kamangar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2039-2048

This paper proposes a novel deep subspace clustering approach which uses convolutional autoencoders to transform input images into new representations lying on a union of linear subspaces. The first contribution of our work is to insert multiple fully-connected linear layers between the encoder layers and their corresponding decoder layers to promote learning more favorable representations for subspace clustering. These connection layers facilitate the feature learning procedure by combining low-level and high-level information for generating multiple sets of self-expressive and informative representations at different levels of the encoder. Moreover, we introduce a novel loss minimization problem which leverages an initial clustering of the samples to effectively fuse the multi-level representations and recover the underlying subspaces more accurately. The loss function is then minimized through an iterative scheme which alternatively updates the network parameters and produces new clusterings of the samples. Experiments on four real-world datasets demonstrate that our approach exhibits superior performance compared to the state-of-the-art methods on most of the subspace clustering problems.

****

VRT-Net: Real-Time Scene Parsing via Variable Resolution Transform

Jogendra Nath Kundu, Gaurav Singh Rajput, Venkatesh Babu RADHAKRISHNAN; Procee dings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2049-2056

Urban scene parsing is a basic requirement for various autonomous navigation systems especially in self-driving. Most of the available approaches employ generic image parsing architectures designed for segmentation of object focused scene captured in indoor setups. However, images captured in car-mounted cameras exhibit an extreme effect of perspective geometry, causing a significant scale disparity between near and farther objects. Recognizing this, we formalize a unique Variable Resolution Transform (VRT) technique motivated from the foveal magnification in human eye. Following this, we design a Fovea Estimation Network (FEN) which is trained to estimate a single most convenient fixation location along with the associated magnification factor, best suited for a given input image. The proposed framework is designed to enable its usage as a wrapper over the available real-time scene parsing models, thereby demonstrating a superior trade-off between speed and quality as compared to the prior state-of-the-arts.

****

RPM-Net: Robust Pixel-Level Matching Networks for Self-Supervised Video Object Segmentation

Youngeun Kim, Seokeon Choi, Hankyeol Lee, Taekyung Kim, Changick Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2057-2065

In this paper, we introduce a self-supervised approach for video object segmentation without human labeled data. Specifically, we present Robust Pixel-level Matching Networks (RPM-Net), a novel deep architecture that matches pixels between adjacent frames, using only color information from unlabeled videos for training. Technically, RPM-Net can be separated into two main modules. The embedding module first projects input images into high dimensional embedding space. Then the matching module with deformable convolution layers matches pixels between reference and target frames based on the embedding features. Unlike previous supervised methods using deformable convolution, our matching module adopts deformable convolution to focus on similar features in spatiotemporally neighboring pixels. We further propose an online updating module to refine the segmentation result by transferring knowledge from the given first frame. Also, we carry out comprehensive experiments on three public datasets (i.e., DAVIS-2017, SegTrack-v2, and Youtube- Objects) and achieve state-of-the-art performance on self-supervised video object segmentation.

****

SINet: Extreme Lightweight Portrait Segmentation Networks with Spatial Squeeze Module and Information Blocking Decoder

Hyojin Park, Lars Sjosund, Youngjoon Yoo, Nicolas Monet, Jihwan Bang, Nojun Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2066-2074

Designing a lightweight and robust portrait segmentation algorithm is an important task for a wide range of face applications. However, the problem has been considered as a subset of the object segmentation and less handled in this field. Obviously, portrait segmentation has its unique requirements. First, because the portrait segmentation is performed in the middle of a whole process, it requires extremely lightweight models. Second, there has not been any public datasets in this domain that contain a sufficient number of images. To solve the first problem, we introduce the new extremely lightweight portrait segmentation model SINet, containing an information blocking decoder and spatial squeeze modules. The information blocking decoder uses confidence estimation to recover local spatial information without spoiling global consistency. The spatial squeeze module uses multiple receptive fields to cope with various sizes of consistency. To tackle the second problem, we propose a simple method to create additional portrait segmentation data, which can improve accuracy. In our qualitative and quantitative analysis on the EG1800 dataset, we show that our method outperforms various existing lightweight models. Our method reduces the number of parameters from 2:1M to 86:9K (around 95.9% reduction), while maintaining the accuracy under an 1% margin from the state-of-the-art method. We also show our model is successfully exe cuted on a real mobile device with 100.6 FPS. In addition, we demonstrate that our method can be used for general semantic segmentation on the Cityscapes dataset. The code and dataset are available in https://github.com/HYOJINPARK/ExtPortraitSeg.

****

Multi-Modal Association based Grouping for Form Structure Extraction

Milan Aggarwal, Mausoom Sarkar, Hiresh Gupta, Balaji Krishnamurthy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2075-2084

Document structure extraction has been a widely researched area for decades. Recent work in this direction has been deep learning-based, mostly focusing on extracting structure using fully convolution NN through semantic segmentation. In this work, we present a novel multi-modal approach for form structure extraction. Given simple elements such as textruns and widgets, we extract higher-order structures such as TextBlocks, Text Fields, Choice Fields, and Choice Groups, which are essential for information collection in forms. To achieve this, we obtain a local image patch around each low-level element (reference) by identifying candidate elements closest to it. We process textual and spatial representation of candidates sequentially through a BiLSTM to obtain context-aware representations and fuse them with image patch features obtained by processing it through a CNN. Subsequently, the sequential decoder takes this fused feature vector to predict the association type between reference and candidates. These predicted associations are utilized to determine larger structures through connected components analysis. Experimental results show the effectiveness of our approach achieving a recall of 90.29%, 73.80%, 83.12%, and 52.72% for the above structures, respectively, outperforming semantic segmentation baselines significantly. We show the efficacy of our method through ablations, comparing it against using individual modalities. We also introduce our new rich human-annotated Forms Dataset.

****

Multiparty Visual Co-Occurrences for Estimating Personality Traits in Group Meetings

Lingyu Zhang, Indrani Bhattacharya, Mallory Morgan, Michael Foley, Christoph Riedl, Brooke Welles, Richard Radke; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2085-2094

Participants' body language during interactions with others in a group meeting can reveal important information about their individual personalities, as well as their contribution to a team. Here, we focus on the automatic extraction of visual features from each person, including her/his facial activity, body movement, and hand position, and how these features co-occur among team members (e.g., how frequently a person moves her/his arms or makes eye contact when she/he is the focus of attention of the group). We correlate these features with user questionnaires to reveal relationships with the "Big Five" personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), as well as with team judgements about the leader and dominant contributor in a conversation. We demonstrate that our algorithms achieve state-of-the-art accuracy with an average of 80% for Big-Five personality trait prediction, potentially enabling integration into automatic group meeting understanding systems.

****

Uncertainty-aware Short-term Motion Prediction of Traffic Actors for Autonomous Driving

Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, NITIN SINGH, Jeff Schneider; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2095-2104

We address one of the crucial aspects necessary for safe and efficient operations of autonomous vehicles, namely predicting future state of traffic actors in the autonomous vehicle's surroundings. We introduce a deep learning-based approach that takes into account a current world state and produces raster images of each actor's vicinity. The rasters are then used as inputs to deep convolutional models to infer future movement of actors while also accounting for and capturing inherent uncertainty of the prediction task. Extensive experiments on real-world data strongly suggest benefits of the proposed approach. Moreover, following successful tests the system was deployed to a fleet of autonomous vehicles.

****

Image identification of Protea species with attributes and subgenus scaling

Peter Thompson, Willie Brink; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2105-2113

The flowering plant genus Protea is a dominant representative for the biodiversity of the Cape Floristic Region in South Africa, and from a conservation point of view important to monitor. The recent surge in popularity of crowd-sourced wildlife monitoring platforms presents challenges and opportunities for automatic image based species identification. We consider the problem of identifying the Protea species in a given image with additional (but optional) attributes linked to the observation, such as location and date. We collect training and test data from a crowd-sourced platform, and find that the Protea identification problem is exacerbated by considerable inter-class similarity, data scarcity, class imbalance, as well as large variations in image quality, composition and background. Our proposed solution consists of three parts. The first part incorporates a variant of multi-region attention into a pretrained convolutional neural network, to focus on the flowerhead in the image. The second part performs coarser-grained classification on subgenera (superclasses) and then rescales the output of the first part. The third part conditions a probabilistic model on the additional attributes associated with the observation. We perform an ablation study on the proposed model and its constituents, and find that all three components together outperform our baselines and all other variants quite significantly.

****

Reverse Variational Autoencoder for Visual Attribute Manipulation and Anomaly Detection

Gauerhof Lydia, Nianlong Gu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2114-2123

In this paper, we introduce the `Reverse Variational Autoencoder" (Reverse-VAE) which is a generative network. On the one hand, visual attributes can be manipulated and combined while generating images. On the other hand, anomalies, meaning deviations from the data space used for training, can be detected. During training the generator network maps samples from stochastic latent vectors to the data space. Meanwhile the encoder network takes these generated images to reconstruct the latent vector. The generator and discriminator are trained adversarially. The discriminator is trained to distinguish between real and generated data. Overall, our model tries to match the joint latent/data-space distribution of the generator and the latent/data-space joint distribution of the encoder by minimizing their Kullback-Leibler divergence. Desired visual attributes of CelebA images are successfully manipulated. The performance of anomaly detection is competitive with state-of-the-art on MNIST and KDD 99 data set.

****

QR-code Reconstruction from Event Data via Optimization in Code Subspace

Jun Nagata, Yusuke Sekikawa, Kosuke Hara, Teppei Suzuki, Yoshimitsu Aoki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2124-2132

We propose an image reconstruction method from event data, assuming the target images belong to a prespecified class like QR codes. Instead of solving the reconstruction problem in the image space, we introduce a code space that covers all the noiseless target class images and solves the reconstruction problem on it. This restriction enormously reduces the number of optimizing parameters and makes the reconstruction problem well posed and robust to noise. We demonstrate fast and robust QR-code scanning in difficult, high-speed scenes with industrial high-speed cameras and other reconstruction methods.

****

Region Pooling with Adaptive Feature Fusion for End-to-End Person Recognition

Vijay Kumar, Anoop Namboodiri, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2133-2142

Current approaches for person recognition train an ensemble of region specific c onvolutional neural networks for representation learning, and then adopt naive fusion strategies to combine their features or predictions during testing. In this paper, we propose an unified end-to-end architecture that generates a complete person representation based on pooling and aggregation of features from multiple body regions. Our network takes a person image and the pre-determined locations of body regions as input, and generates common feature maps that are shared across all the regions. Multiple features corresponding to different regions are then pooled and combined with an aggregation block, where the adaptive weights required for aggregation are obtained through an attention mechanism. Evaluations on three person recognition datasets - PIPA, Soccer and Hannah show that a single model trained end-to-end is computationally faster, requires fewer parameters and achieves improved performance over separately trained models.

****

Event-based Star Tracking via Multiresolution Progressive Hough Transforms

Samya Bagchi, Tat-Jun Chin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2143-2152

Star trackers are state-of-the-art attitude estimation devices which function by recognising and tracking star patterns. Most commercial star trackers use conventional optical sensors. A recent alternative is to use event sensors, which could enable more energy-efficient and faster star trackers. However, this demands new algorithms that can efficiently cope with high-speed asynchronous data, and are feasible on resource-constrained computing platforms. To this end, we propose an event-based processing approach for star tracking. Our technique operates on the event stream from a star field, by using multiresolution Hough Transforms to time-progressively integrate event data and produce accurate relative rotations. Optimisation via rotation averaging is then used to fuse the relative rotations and jointly refine the absolute orientations. Our technique is designed to be feasible for asynchronous operation on standard hardware. Moreover, compared to state-of-the-art event-based motion estimation schemes, our technique is much more efficient and accurate.

****

Toward Explainable Fashion Recommendation

Pongsate Tangseng, Takayuki Okatani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2153-2162

Many studies have been conducted so far to build systems for recommending fashion items and outfits. Although they achieve good performances in their respective tasks, most of them cannot explain their judgments to the users, which compromises their usefulness. Toward explainable fashion recommendation, this study proposes a system that is able not only to provide a goodness score for an outfit but also to explain the score by providing reason behind it. For this purpose, we propose a method for quantifying how influential each feature of each item is to the score. Using this influence value, we can identify which item and what feature make the outfit good or bad. We represent the image of each item with a combination of human-interpretable features, and thereby the identification of the most influential item-feature pair gives useful explanation of the output score. To evaluate the performance of this approach, we design an experiment that can be performed without human annotation; we replace a single item-feature pair in an outfit so that the score will decrease, and then we test if the proposed method can detect the replaced item-feature pair correctly using the above influence values. The experimental results show that the proposed method can accurately detect bad items in outfits lowering their scores.

****

Self-Contained Stylization via Steganography for Reverse and Serial Style Transfer

Hung-Yu Chen, I-Sheng Fang, Chia-Ming Cheng, Wei-Chen Chiu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2163-2171

Style transfer has been widely applied to give real-world images a new artistic look. However, given a stylized image, the attempts to use typical style transfer methods for de-stylization or transferring it again into another style usually lead to artifacts or undesired results. We realize that these issues are originated from the content inconsistency between the original image and its stylized output. Therefore, in this paper we advance to keep the content information of the input image during the process of style transfer by the power of steganography, with two approaches proposed: a two-stage model and an end-to-end model. We conduct extensive experiments to successfully verify the capacity of our models, in which both of them are able to not only generate stylized images of quality comparable with the ones produced by typical style transfer methods, but also effectively eliminate the artifacts introduced in reconstructing original input from a stylized image as well as performing multiple times of style transfer in series.

****

Adversarial Discriminative Attention for Robust Anomaly Detection

Daiki Kimura, Subhajit Chaudhury, Minori Narita, Asim Munawar, Ryuki Tachibana; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2172-2181

Existing methods for visual anomaly detection predominantly rely on global level pixel comparisons for anomaly score computation without emphasizing on unique local features. However, images from real-world applications are susceptible to unwanted noise and distractions, that might jeopardize the robustness of such anomaly score. To alleviate this problem, we propose a self-supervised masking method that specifically focuses on discriminative parts of images to enable robust anomaly detection. Our experiments reveal that discriminator's class activation map in adversarial training evolves in three stages and finally fixates on the foreground location in the images. Using this property of the activation map, we construct a mask that suppresses spurious signals from the background thus enabling robust anomaly detection by focusing on local discriminative attributes. Additionally, our method can further improve the accuracy by learning a semi-supervised discriminative classifier in cases where a few samples from anomaly classes are available during the training. Experimental evaluations on four different types of datasets demonstrate that our method outperforms previous state-of-the-art methods for each condition and in all domains.

****

SieveNet: A Unified Framework for Robust Image-Based Virtual Try-On

Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Balaji Krishnamurthy, Abhijeet Halwai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2182-2190

Image-based virtual try-on for fashion has attracted considerable attention recently. The task requires trying on the desired clothing item on a target model. An efficient framework for this is composed of 2 stages: (1) warping (transforming) the try-on cloth to align with the pose and shape of the target model, and (2) a texture transfer module to seamlessly integrate the warped try-on cloth onto the target model image. Existing methods suffer from artifacts and distortions in their try-on output. In this work, we present SieveNet, a framework for robust image-based virtual try-on. Firstly, we introduce a multi-stage coarse-to-fine warping network to better model fine-grained intricacies in try-on clothing item and train it with a novel perceptual geometric matching loss. Next, we introduce a try-on cloth conditioned segmentation mask prior to improve the texture transfer network. Finally, we also introduce a dueling triplet strategy for training the texture transfer network which further improves the quality of the generated try-on result. We present extensive qualitative and quantitative evaluations on each component of the proposed pipeline and show significant performance improvements against existing state-of-the-art methods.

****

A Flexible Selection Scheme for Minimum-Effort Transfer Learning

Amelie Royer, Christoph Lampert; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2191-2200

Fine-tuning is a popular way of exploiting knowledge contained in a pre-trained convolutional network for a new visual recognition task. However, the orthogonal setting of transferring knowledge from a pretrained network from a visually dif ferent yet semantically close source is rarely considered: This commonly happens with real-life data which is not necessarily as clean as the training source (noise, ge- ometric transformations, different modalities, etc.). To tackle such scenarios, we introduce a new, generalized form of fine-tuning, called flex-tuning, in which any individual unit (e.g. layer) of a network can be tuned, and the most promising one is chosen automatically. In order to make the method appealing for practical use, we propose two lightweight and faster selection procedures that prove to be good approximations in practice. We study these selection criteria empirically across a variety of domain shifts and data scarcity scenarios, and show that fine-tuning individual units, despite its simplicity, yields very good results as an adaptation technique. As it turns out, in contrast to common practice, rather than the last fully-connected unit it is best to tune an intermediate or early one in many domain-shift scenarios, which is accurately detected by flex-tuning.

****

Multi-Label Visual Feature Learning with Attentional Aggregation

Ziqiao Guan, Kevin Yager, Dantong Yu, Hong Qin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2201-2209

Today convolutional neural networks (CNNs) have reached out to specialized applications in science communities that otherwise would not be adequately tackled. In this paper, we systematically study a multi-label annotation problem of x-ray scattering images in material science. For this application, we tackle an open challenge with training CNNs --- identifying weak scattered patterns with diffuse background interference, which is common in scientific imaging. We articulate an Attentional Aggregation Module (AAM) to enhance feature representations. First, we reweight and highlight important features in the images using data-driven attention maps. We decompose the attention maps into channel and spatial attention components. In the spatial attention component, we design a mechanism to generate multiple spatial attention maps tailored for diversified multi-label learning. Then, we condense the enhanced local features into non-local representations by performing feature aggregation. Both attention and aggregation are designed as network layers with learnable parameters so that CNN training remains fluidly end-to-end, and we apply it in-network a few times so that the feature enhancement is multi-scale. We conduct extensive experiments on CNN training and testing, as well as transfer learning, and empirical studies confirm that our method enhances the discriminative power of visual features of scientific imaging.

****

Instance Segmentation for the Quantification of Microplastic Fiber Images

Viktor Wegmayr, Aytunc Sahin, Bjorn Saemundsson, Joachim Buhmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2210-2217

Microplastics pollution has been recognized as a serious environmental concern, with serious research efforts underway to determine primary causes. Experiments typically generate bright-field images of microplastic fibers that are filtered from water. Environmental decision making in process engineering critically relies on accurate quantification of microplastic fibers in these images. To satisfy the required standards, images are often analyzed manually, resulting in a highly tedious process, with thousands of fiber instances per image. While the shape of individual fibers is relatively simple, it is difficult to separate them in highly crowded scenes with significant overlap. We propose a fiber instance detection pipeline, which decomposes the fiber detection and segmentation into manageable subproblems. Well separated instances are identified with robust image processing techniques, such as adaptive thresholding, and morphological skeleton analysis, while tangled fibers are separated by an algorithm based on deep pixel embeddings. Moreover, we present a modified Intersection-over- Union metric as a more appropriate similarity metric for elongated shapes. Our approach improves significantly on out-of-sample data, in particular for difficult cases of intersecting fibers.

****

Charting the Right Manifold: Manifold Mixup for Few-shot Learning

Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2218-2227

Few-shot learning algorithms aim to learn model parameters capable of adapting to unseen classes with the help of only a few labeled examples. A recent regularization technique - Manifold Mixup focuses on learning a general-purpose representation, robust to small changes in the data distribution. Since the goal of few-shot learning is closely linked to robust representation learning, we study Manifold Mixup in this problem setting. Self-supervised learning is another technique that learns semantically meaningful features, using only the inherent structure of the data. This work investigates the role of learning relevant feature manifold for few-shot tasks using self-supervision and regularization techniques. We observe that regularizing the feature manifold, enriched via self-supervised techniques, with Manifold Mixup significantly improves few-shot learning performance. We show that our proposed method S2M2 beats the current state-of-the-art accuracy on standard few-shot learning datasets like CIFAR-FS, CUB, mini-ImageNet and tiered-ImageNet by 3-8 %. Through extensive experimentation, we show that the features learned using our approach generalize to complex few-shot evaluation tasks, cross-domain scenarios and are robust against slight changes to data distribution.

****

Measuring the Utilization of Public Open Spaces by Deep Learning: a Benchmark Study at the Detroit Riverfront

Peng Sun, Rui Hou, Jerome Lynch; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2228-2237

Physical activities and social interactions are essential activities that ensure a healthy lifestyle. Public open spaces (POS), such as parks, plazas and greenways, are key environments that encourage those activities. To evaluate a POS, there is a need to study how humans use the facilities within it. However, traditional approaches to studying use of POS are manual and therefore time and labor intensive. They also may only provide qualitative insights. It is appealing to make use of surveillance cameras and to extract user-related information through computer vision. This paper proposes a proof-of-concept deep learning computer vision framework for measuring human activities quantitatively in POS and demonstrates a case study of the proposed framework using the Detroit Riverfront Conservancy (DRFC) surveillance camera network. A custom image dataset is presented to train the framework; the dataset includes 7826 fully annotated images collected from 18 cameras across the DRFC park space under various illumination conditions. Dataset analysis is also provided as well as a baseline model for one-step user localization and activity recognition. The mAP results are 77.5% for pedestrian detection and 81.6% for cyclist detection. Behavioral maps are autonomously generated by the framework to locate different POS users and the average error for behavioral localization is within 10 cm.

****

Erasing Scene Text with Weak Supervision

Jan Zdenek, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2238-2246

Scene text erasing is a task of removing text from natural scene images, which has been gaining attention in recent years. The main motivation is to conceal private information such as license plate numbers, and house nameplates that can appear in images. In this work, we propose a method for scene text erasing that approaches the problem as a general inpainting task. In contrast to previous methods, which require pairs of original images containing text and images from which the text has been removed, our method does not need corresponding image pairs for training. We use a separately trained scene text detector and an inpainting network. The scene text detector predicts segmentation maps of text instances which are then used as masks for the inpainting network. The network for inpainting, trained on a large-scale image dataset, fills in masked out regions in an input image and generates a final image in which the original text is no longer pres ent. The results show that our method is able to successfully remove text and fill in the created holes to produce natural-looking images.

****

Scalable Detection of Offensive and Non-compliant Content / Logo in Product Images

Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, Shie Mannor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2247-2256

In e-commerce, product content, especially product images have a significant influence on a customer's journey from product discovery to evaluation and finally, purchase decision. Since many e-commerce retailers sell items from other third-party marketplace sellers besides their own, the content published by both internal and external content creators needs to be monitored and enriched, wherever possible. Despite guidelines and warnings, product listings that contain offensive and non-compliant images continue to enter catalogs. Offensive and non-compliant content can include a wide range of objects, logos, and banners conveying violent, sexually explicit, racist, or promotional messages. Such images can severely damage the customer experience, lead to legal issues, and erode the company brand. In this paper, we present a computer vision driven offensive and non-compliant image detection system for extremely large image datasets. This paper delves into the unique challenges of applying deep learning to real-world product image data from retail world. We demonstrate how we resolve a number of technical challenges such as lack of training data, severe class imbalance, fine-grained class definitions etc. using a number of practical yet unique technical strategies. Our system combines state-of-the-art image classification and object detection techniques with budgeted crowdsourcing to develop a solution customized for a massive, diverse, and constantly evolving product catalog.

****

Very Power Efficient Neural Time-of-Flight

Yan Chen, Jimmy Ren, Xuanye Cheng, Keyuan Qian, Luyang Wang, Jinwei Gu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2257-2266

Time-of-Flight (ToF) cameras require active illumination to obtain depth information thus the power of illumination directly affects the performance of ToF cameras. Traditional ToF imaging algorithms are very sensitive to illumination and the depth accuracy degenerates rapidly with the power of it. Therefore, the design of a power efficient ToF camera always creates a painful dilemma for the illumination and the performance trade-off. In this paper, we show that despite the weak signals in many areas under extreme short exposure setting, these signals as a whole can be well utilized through a learning process which directly translates the weak and noisy ToF camera raw to depth map. This creates an opportunity to tackle the aforementioned dilemma and make a very power efficient ToF camera possible. To enable the learning, we collect a comprehensive dataset under a variety of scenes and photographic conditions by a specialized ToF camera. Experiments show that our method is able to robustly process ToF camera raw with the exposure time of one order of magnitude shorter than that used in conventional ToF cameras. In addition to evaluating our approach both quantitatively and qualitatively, we also discuss its implication to designing the next generation power efficient ToF cameras.

****

Semantic Consistency and Identity Mapping Multi-Component Generative Adversarial Network for Person Re-Identification

Amena Khatun, SIMON DENMAN, Sridha Sridharan, Clinton Fookes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2267-2276

In a real world environment, person re-identification (Re-ID) is a challenging task due to variations in lighting conditions, viewing angles, pose and occlusions. Despite recent performance gains, current person Re-ID algorithms still suffer heavily when encountering these variations. To address this problem, we propos e a semantic consistency and identity mapping multi-component generative adversarial network (SC-IMGAN) which provides style adaptation from one to many domains. To ensure that transformed images are as realistic as possible, we propose novel identity mapping and semantic consistency losses to maintain identity across the diverse domains. For the Re-ID task, we propose a joint verification-identification quartet network which is trained with generated and real images, followed by an effective quartet loss for verification. Our proposed method outperforms state-of-the-art techniques on six challenging person Re-ID datasets: CUHK01, CUHK03, VIPeR, PRID2011, iLIDS and Market-1501.

****

Structured Compression of Deep Neural Networks with Debiased Elastic Group LASSO
Oyebade Oyedotun, Djamila Aouada, Bjorn Ottersten; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2277-2286

State-of-the-art Deep Neural Networks (DNNs) are typically too cumbersome to be practically useful in portable electronic devices. As such, several works pursue model compression that seeks to drastically reduce computational memory footprints, FLOPS and memory for storage. Many of these works achieve unstructured compression, where the compressed models are not directly useful since dedicated hardware and specialized algorithms are required for storage of sparse weights and fast sparse matrix-vector multiplication respectively. In this paper, we propose structured compression of large DNNs using debiased elastic group LASSO (DEGL), which is motivated by different interesting characteristics of the individual components. That is, where group LASSO penalty enforces structured sparsity, l2-norm penalty promotes features grouping, and debiasing disentangles sparsity and shrinkage effects of group LASSO. We perform extensive experiments by applying DEGL to different DNN architectures including LeNet, VGG, AlexNet and ResNet on MNIST, CIFAR-10, CIFAR-100 and ImageNet datasets. Furthermore, we validate the effectiveness of our proposal on domain adaptation using Oxford-102 flower species and Food-5K datasets. Results show that DEGL can compress DNNs by several folds with small or no loss of performance. Particularly, DEGL outperforms conventional group LASSO and several other state-of-the-art methods that perform structured compression.

****

Plug-and-Play Rescaling Based Crowd Counting in Static Images
Usman Sajid, Guanghui Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2287-2296

Crowd counting is a challenging problem especially in the presence of huge crowd diversity across images and complex cluttered crowd-like background regions, where most previous approaches do not generalize well and consequently produce either huge crowd underestimation or overestimation. To address these challenges, we propose a new image patch rescaling module (PRM) and three independent PRM employed crowd counting methods. The proposed frameworks use the PRM module to rescale the image regions (patches) that require special treatment, whereas the classification process helps in recognizing and discarding any cluttered crowd-like background regions which may result in overestimation. Experiments on three standard benchmarks and cross-dataset evaluation show that our approach outperforms the state-of-the-art models in the RMSE evaluation metric with an improvement up to 10:4%, and possesses superior generalization ability to new datasets.

****

Looking Ahead: Anticipating Pedestrians Crossing with Future Frames Prediction
Mohamed Chaabane, Ameni Trabelsi, Nathaniel Blanchard, Ross Beveridge; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2297-2306

In this paper, we present an end-to-end future-prediction model that focuses on pedestrian safety. Specifically, our model uses previous video frames, recorded from the perspective of the vehicle, to predict if a pedestrian will cross in front of the vehicle. The long term goal of this work is to design a fully autonomous system that acts and reacts as a defensive human driver would --- predicting future events and reacting to mitigate risk. We focus on pedestrian-vehicle int eractions because of the high risk of harm to the pedestrian if their actions are miss-predicted. Our end-to-end model consists of two stages: the first stage is an encoder-decoder network that learns to predict future video frames. The second stage is a deep spatio-temporal network that utilizes the predicted frames of the first stage to predict the pedestrian's future action. Our system achieves state-of-the-art accuracy on pedestrian behavior prediction and future frames prediction on the Joint Attention for Autonomous Driving (JAAD) dataset.

****

Post-Mortem Iris Recognition Resistant to Biological Eye Decay Processes

Mateusz Trokielewicz, Adam Czajka, Piotr Maciejewicz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2307-2315

This paper proposes an end-to-end iris recognition method designed specifically for post-mortem samples, and thus serving as a perfect application for iris biometrics in forensics. To our knowledge, it is the first method specific for verification of iris samples acquired after demise. We have fine-tuned a convolutional neural network-based segmentation model with a large set of diversified iris data (including post-mortem and diseased eyes), and combined Gabor kernels with newly designed, iris-specific kernels learnt by Siamese networks. The resulting method significantly outperforms the existing off-the-shelf iris recognition methods (both academic and commercial) on the newly collected database of post-mortem iris images and for all available time horizons since death. We make all models and the method itself available along with this paper.

****

Fungi Recognition: A Practical Use Case

Milan Sulc, Lukas Picek, Jiri Matas, Thomas Jeppesen, Jacob Heilmann-Clausen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2316-2324

The paper presents a system for visual recognition of 1394 fungi species based on deep convolutional neural networks and its deployment in a citizen-science project. The system allows users to automatically identify observed specimens, while providing valuable data to biologists and computer vision researchers. The underlying classification method scored first in the FGVCx Fungi Classification Kaggle competition organized in connection with the Fine-Grained Visual Categorization (FGVC) workshop at CVPR 2018. We describe our winning submission and evaluate all technicalities that increased the recognition scores, and discuss the issues related to deployment of the system via the web- and mobile- interfaces.

****

CompressNet: Generative Compression at Extremely Low Bitrates

Suraj Kiran Raman, Aditya Ramesh, Vijayakrishna Naganoor, Shubham Dash, Giridharan Kumaravelu, Honglak Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2325-2333

Compressing images at extremely low bitrates (< 0.1 bpp) has always been a challenging task as the quality of reconstruction significantly reduces due to the strongly imposing constraint on the number of bits allocated for the compressed data. With the increasing need to transfer large amounts of images with limited bandwidth, compressing images to very low sizes is a crucial task. However, the existing methods are not effective at extremely low bitrates. To address this need we propose a novel network called CompressNet which augments a Stacked Autoencoder with a Switch Prediction Network (SAE-SPN). This helps in the reconstruction of visually pleasing images at these low bitrates (< 0.1 bpp). We benchmark the performance of our proposed method on the Cityscapes dataset, evaluating over different metrics at very low bitrates showing that our method outperforms the other state-of-the-art. In particular, at a bitrate of 0.07, CompressNet achieves 22% lower Perceptual Loss and 55% lower Frechet Inception Distance (FID) compared to the deep learning SOTA methods.

****

A GAN-based Tunable Image Compression System

Lirong Wu, Kejie Huang, Haibin Shen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2334-2342

The method of importance map has been widely adopted in DNN-based lossy image compression to achieve bit allocation according to the importance of image contents. However, insufficient allocation of bits in non-important regions often leads to severe distortion at low bpp (bits per pixel), which hampers the development of efficient content-weighted image compression systems. This paper rethinks content-based compression by using Generative Adversarial Network (GAN) to reconstruct the non-important regions. Moreover, multiscale pyramid decomposition is applied to both the encoder and the discriminator to achieve global compression of high-resolution images. A tunable compression scheme is also proposed in this paper to compress an image to any specific compression ratio without retraining the model. The experimental results show that our proposed method improves MS-SSIM by more than 10.3% compared to the recently reported GAN-based method to achieve the same low bpp (0.05) on the Kodak dataset.

****

Going Much Wider with Deep Networks for Image Super-Resolution
Vikram Singh, Keerthan Ramnath, Subrahmanyam Arunachalam, Anurag Mittal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2343-2354

Divide and Conquer is a well-established approach in the literature that has efficiently solved a variety of problems. However, it is yet to be explored in full in solving image super-resolution. To predict a sharp up-sampled image, this work proposes a divide and conquer approach based wide and deep network (WDN) that divides the 4x up-sampling problem into 32 disjoint subproblems that can be solved simultaneously and independently of each other. Half of these subproblems deal with predicting the overall features of the high-resolution image, while the remaining are exclusively for predicting the finer details. Additionally, a technique that is found to be more effective in calibrating the pixel intensities has been proposed. Results obtained on multiple datasets demonstrate the improved performance of the proposed wide and deep network over state-of-the-art methods.

****

Scale-aware Conditional Generative Adversarial Network for Image Dehazing
Prasen Sharma, Priyankar Jain, Arijit Sur; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2355-2365

Outdoor images are often deteriorated due to the presence of haze in the atmosphere. Conventionally, the single image dehazing problem aims to restore the haze-free image. Previous successful approaches have utilized various hand-crafted features/priors. However, such images suffer from color degradation and halo artifacts. By way of analysis, these artifacts, in general, prevail around the regions with high-intensity variation, such as edgy structures. This finding inspires us to consider the Laplacians of Gaussian (LoG) of the images which exceptionally retains this information, to solve the problem of single image haze removal. In this line of thought, we present an end-to-end model that learns to remove the haze based on the per-pixel difference between LoGs of the dehazed and original haze-free images. The optimization of the proposed network is further enhanced by using the adversarial training and perceptual loss function. The proposed method has been appraised on Synthetic Objective Testing Set (SOTS) and benchmark realworld hazy images using 16 image quality measures. Based on the Color Difference (CIEDE 2000), an improvement of   15.89% has been observed over the state-of-the-art method, Yang et al. [50]. An ablation study has been presented at the end to illustrate the improvements achieved by various modules of the proposed network.

****

DCIL: Deep Contextual Internal Learning for Image Restoration and Image Retargeting
Indra Deep Mastan, Shanmuganathan Raman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2366-2375

Recently, there is a vast interest in developing methods which are independent of the training samples such as deep image prior, zero-shot learning, and internal learning. The methods above are based on the common goal of maximizing image features learning from a single image despite inherent technical diversity. In th is work, we bridge the gap between the various unsupervised approaches above and propose a general framework for image restoration and image retargeting. We use contextual feature learning and internal learning to improvise the structure similarity between the source and the target images. We perform image resize application in the following setups: classical image resize using super-resolution, a challenging image resize where the low-resolution image contains noise, and content-aware image resize using image retargeting. We also provide comparisons to the relevant state-of-the-art methods.

****

DAVID: Dual-Attentional Video Deblurring

Junru Wu, Xiang Yu, Ding Liu, Manmohan Chandraker, Zhangyang Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2376-2385

Blind video deblurring restores sharp frames from a blurry sequence without any prior. It is a challenging task because the blur due to camera shake, object movement and defocusing is heterogeneous in both temporal and spatial dimensions. Traditional methods train on datasets synthesized with a single level of blur, and thus do not generalize well across levels of blurriness. To address this challenge, we propose a dual attention mechanism to dynamically aggregate temporal cues for deblurring with an end-to-end trainable network structure. Specifically, an internal attention module adaptively selects the optimal temporal scales for restoring the sharp center frame. An external attention module adaptively aggregates and refines multiple sharp frame estimates, from several internal attention modules designed for different blur levels. To train and evaluate on more diverse blur severity levels, we propose a Challenging DVD dataset generated from the raw DVD video set by pooling frames with different temporal windows. Our framework achieves consistently better performance on this more challenging dataset while obtaining strongly competitive results on the original DVD benchmark. Extensive ablative studies and qualitative visualizations further demonstrate the advantage of our method in handling real video blur.

****

360 Panorama Synthesis from a Sparse Set of Images with Unknown Field of View

Julius Surya Sumantri, In Kyu Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2386-2395

360 images represent scenes captured in all possible viewing directions and enable viewers to navigate freely around the scene thereby providing an immersive experience. Conversely, conventional images represent scenes in a single viewing direction with a small or limited field of view (FOV). As a result, only certain parts of the scenes are observed, and valuable information about the surroundings is lost. In this paper, a learning-based approach that reconstructs the scene in 360 x 180 from a sparse set of conventional images (typically 4 images) is proposed. The proposed approach first estimates the FOV of input images relative to the panorama. The estimated FOV is then used as the prior for synthesizing a high-resolution 360 panoramic output. The proposed method overcomes the difficulty of learning-based approach in synthesizing high resolution images (up to 512x1024). Experimental results demonstrate that the proposed method produces 360 panorama with reasonable quality. Results also show that the proposed method outperforms the alternative method and can be generalized for non-panoramic scenes and images captured by a smartphone camera.

****

From Image to Video Face Inpainting: Spatial-Temporal Nested GAN (STN-GAN) for Usability Recovery

Yifan Wu, Vivek Singh, Ankur Kapoor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2396-2405

In this paper, we propose to use constrained inpainting methods to recover usability of corrupted images. Here we focus on the example of face images that are masked for privacy protection but complete images are required for further algorithm development. The task is tackled in a progressive manner: 1) the generated images should look realistic; 2) the generated images must satisfy spatial constraints, if available; 3) when applied to video data, temporal consistency should be retained. We first present a spatial inpainting framework to synthesize face images which can incorporate spatial constraints, provided as positions of facial markers and show that it outperforms state-of-the-art methods. Next, we propose Spatial-Temporal Nested GAN (STN-GAN) to adapt image inpainting framework, trained on 200k images, to video data by incorporating temporal information using residual blocks. Experiments on multiple public datasets show STN-GAN attains spatio-temporal consistency effectively and efficiently. Furthermore, we show that spatial constraints can be perturbed to obtain different inpainted results from a single source.

****

Variational Image Deraining

Yingjun Du, Jun Xu, Qiang Qiu, Xiantong Zhen, Lei Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2406-2415

Images captured in severe weather such as rain and snow significantly degrade the accuracy of vision systems, e.g., for outdoor video surveillance or autonomous driving. Image deraining is a critical yet highly challenging task, due to the fact that rain density varies across spatial locations, while the distribution patterns simultaneously vary across color channels. In this paper, we propose a variational image deraining (VID) method by formulating image deraining in a conditional variational auto-encoder framework. To achieve adaptive deraining to spatial rain density, we generate a density estimation map for each color channel, which can largely avoid over and under deraining. In addition, to address cross-channel variations, we conduct channel-wise deraining, motivated by our observation that bright pixels do not tend to remain bright after deraining unless there color channels are handled separately. Experimental results show that the proposed deraining method achieves superior performance on both synthesized and real rainy images, surpassing previous state-of-the-art methods by large margins. The code will be publicly released.

****

End-To-End Trainable Video Super-Resolution Based on a New Mechanism for Implicit Motion Estimation and Compensation

Xiaohong Liu, Lingshi Kong, Yang Zhou, Jiying Zhao, Jun Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2416-2425

Video super-resolution aims at generating a high-resolution video from its low-resolution counterpart. With the rapid rise of deep learning, many recently proposed video super-resolution methods use convolutional neural networks in conjunction with explicit motion compensation to capitalize on statistical dependencies within and across low-resolution frames. Two common issues of such methods are noteworthy. Firstly, the quality of the final reconstructed HR video is often very sensitive to the accuracy of motion estimation. Secondly, the warp grid needed for motion compensation, which is specified by the two flow maps delineating pixel displacements in horizontal and vertical directions, tends to introduce additional errors and jeopardize the temporal consistency across video frames. To address these issues, we propose a novel dynamic local filter network to perform implicit motion estimation and compensation by employing, via locally connected layers, sample-specific and position-specific dynamic local filters that are tailored to the target pixels. We also propose a global refinement network based on ResBlock and autoencoder structures to exploit non-local correlations and enhance the spatial consistency of super-resolved frames. The experimental results demonstrate that the proposed method outperforms the state-of-the-art, and validate its strength in terms of local transformation handling, temporal consistency as well as edge sharpness.

****

Identifying Recurring Patterns with Deep Neural Networks for Natural Image Denoising

Zhihao Xia, Ayan Chakrabarti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2426-2434

Image denoising methods must effectively model, implicitly or explicitly, the va st diversity of patterns and textures that occur in natural images. This is challenging, even for modern methods that leverage deep neural networks trained to regress to clean images from noisy inputs. One recourse is to rely on "internal" image statistics, by searching for similar patterns within the input image itself. In this work, we propose a new method for natural image denoising that trains a deep neural network to determine whether patches in a noisy image input share common underlying patterns. Given a pair of noisy patches, our network predicts whether different sub-band coefficients of the original noise-free patches are similar. The denoising algorithm then aggregates matched coefficients to obtain an initial estimate of the clean image. Finally, this estimate is provided as input, along with the original noisy image, to a standard regression-based denoising network. Experiments show that our method achieves state-of-the-art color image denoising performance, including with a blind version that trains a common model for a range of noise levels, and does not require knowledge of level of noise in an input image. Our approach also has a distinct advantage when training with limited amounts of training data.

****

Characteristic Regularisation for Super-Resolving Face Images
Zhiyi Cheng, Xiatian Zhu, Shaogang Gong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2435-2444

Existing facial image super-resolution (SR) methods focus mostly on improving "artificially down-sampled" low-resolution (LR) imagery. Such SR models, although strong at handling artificial LR images, often suffer from significant performance drop on genuine LR test data. Previous unsupervised domain adaptation (UDA) methods address this issue by training a model using unpaired genuine LR and HR data as well as cycle consistency loss formulation. However, this renders the model overstretched with two tasks: consistifying the visual characteristics and enhancing the image resolution. Importantly, this makes the end-to-end model training ineffective due to the difficulty of back-propagating gradients through two concatenated CNNs. To solve this problem, we formulate a method that joins the advantages of conventional SR and UDA models. Specifically, we separate and control the optimisations for characteristics consistifying and image super-resolving by introducing Characteristic Regularisation (CR) between them. This task split makes the model training more effective and computationally tractable. Extensive evaluations demonstrate the performance superiority of our method over state-of-the-art SR and UDA models on both genuine and artificial LR facial imagery data.

****

ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution
Patricia Vitoria, Lara Raad, Coloma Ballester; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2445-2454

The colorization of grayscale images is an ill-posed problem, with multiple correct solutions. In this paper, we propose an adversarial learning colorization approach coupled with semantic information. A generative network is used to infer the chromaticity of a given grayscale image conditioned to semantic clues. This network is framed in an adversarial model that learns to colorize by incorporating perceptual and semantic understanding of color and class distributions. The model is trained via a fully self-supervised strategy. Qualitative and quantitative results show the capacity of the proposed method to colorize images in a realistic way achieving state-of-the-art results.

****

Image denoising via K-SVD with primal-dual active set algorithm
Quan Xiao, Canhong Wen, Zirui Yan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2455-2463

K-SVD algorithm has been successfully applied to image denoising tasks dozens of years but the big bottleneck in speed and accuracy still needs attention to break. For the sparse coding stage in K-SVD, which involves l0 constraint, prevailing methods usually seek approximate solutions greedily but are less effective once the noise level is high. The alternative l1 optimization is proved to be powerful than l0, however, the time consumption prevents it from the implementation.

In this paper, we propose a new K-SVD framework called K-SVDp by applying the Primal-dual active set (PDAS) algorithm to it. Different from the greedy algorithms based K-SVD, the K-SVDp algorithm develops a selection strategy motivated by KKT (Karush-Kuhn-Tucker) condition and yields to an efficient update in the sparse coding stage. Since the K-SVDp algorithm seeks for an equivalent solution to the dual problem iteratively with simple explicit expression in this denoising problem, speed and quality of denoising can be reached simultaneously. Experiments are carried out and demonstrate the comparable denoising performance of our K-SVDp with state-of-the-art methods.

****

Proposal-free Temporal Moment Localization of a Natural-Language Query in Video using Guided Attention

Cristian Rodriguez, Edison Marrese-Taylor, Fatemeh Sadat Saleh, HONGDONG LI, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2464-2473

This paper studies the problem of temporal moment localization in a long untrimmed video using natural language as the query. Given an untrimmed video and a query sentence, the goal is to determine the start and end of the relevant visual moment in the video that corresponds to the query sentence. While most previous works have tackled this by a propose-and-rank approach, we introduce a more efficient, end-to-end trainable, and proposal-free approach that is built upon three key components: a dynamic filter which adaptively transfers language information to visual domain attention map, a new loss function to guide the model to attend the most relevant part of the video, and soft labels to cope with annotation uncertainties. Our method is evaluated on three standard benchmark datasets, Charades-STA, TACoS and ActivityNet-Captions. Experimental results show our method outperforms state-of-the-art methods on these datasets, confirming the effectiveness of the method. We believe the proposed dynamic filter-based guided attention mechanism will prove valuable for other vision and language tasks as well.

****

Improved Embeddings with Easy Positive Triplet Mining

Hong Xuan, Abby Stylianou, Robert Pless; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2474-2482

Deep metric learning seeks to define an embedding where semantically similar images are embedded to nearby locations, and semantically dissimilar images are embedded to distant locations. Substantial work has focused on loss functions and strategies to learn these embeddings by pushing images from the same class as close together in the embedding space as possible. In this paper, we propose an alternative, loosened embedding strategy that requires the embedding function only map each training image to the most similar examples from the same class, an approach we call "Easy Positive" mining. We provide a collection of experiments and visualizations that highlight that this Easy Positive mining leads to embeddings that are more flexible and generalize better to new unseen data. This simple mining strategy yields recall performance that exceeds state of the art approaches (including those with complicated loss functions and ensemble methods) on image retrieval datasets including CUB, Stanford Online Products, In-Shop Clothes and Hotels-50K.ositive mining leads to embeddings that are more flexibly and generalize better to new data.

****

Learning Discriminative and Generalizable Representations by Spatial-Channel Partition for Person Re-Identification

Hao Chen, Benoit Lagadec, Francois Bremond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2483-2492

In Person Re-Identification (Re-ID) task, combining local and global features is a common strategy to overcome missing key parts and misalignment on models based only on global features. Using this combination, neural networks yield impressive performance in Re-ID task. Previous part-based models mainly focus on spatial partition strategies. Recently, operations on channel information, such as Group Normalization and Channel Attention, have brought significant progress to various visual tasks. However, channel partition has not drawn much attention in Pe rson Re-ID. In this paper, we conduct a study to exploit the potential of channel partition in Re-ID task. Based on this study, we propose an end-to-end Spatial and Channel partition Representation network (SCR) in order to better exploit both spatial and channel information. Experiments conducted on three mainstream image-based evaluation protocols including Market-1501, DukeMTMC-ReID and CUHK03 and one video-based evaluation protocol MARS validate the performance of our model, which outperforms previous state-of-the-art in both single and cross domain Re-ID tasks.

****

Deep Position-Aware Hashing for Semantic Continuous Image Retrieval

Ruikui Wang, Ruiping Wang, Shishi Qiao, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2493-2502

Preserving the semantic similarity is one of the most important goals of hashing. Most existing deep hashing methods employ pairs or triplets of samples in training stage, which only consider the semantic similarity within a mini-batch and depict the local positional relationship in Hamming space, leading to intermittent semantic similarity preservation. In this paper, we propose Deep Position-Aware Hashing (DPAH) to ensure continuous semantic similarity in Hamming space by modeling global positional relationship. Specifically, we introduce a set of learnable class centers as the global proxies to represent the global information and generate discriminative binary codes by constraining the distance between data points and class centers. In addition, in order to reduce the information loss caused by relaxing the binary codes to real-values in optimization, we propose kurtosis loss (KT loss) to handle the distribution of real-valued features before thresholding to be double-peak, and then enable the real-valued features to be more binary-like. Comprehensive experiments on three datasets show that our DPAH outperforms state-of-the-art methods.

****

Spatial-Content Image Search in Complex Scenes

Jin Ma, Shanmin Pang, Bo Yang, Jihua Zhu, Yaochen Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2503-2511

Although the topic of image search has been heavily studied in the last two decades, many works have focused on either instance-level retrieval or semantic-level retrieval. In this work, we develop a novel visually similar spatial-semantic method, namely spatial-content image search, to search images that not only share the same spatial-semantics but also enjoy visual consistency as the query image in complex scenes. We achieve the goal by capturing spatial-semantic concepts as well as the visual representation of each concept contained in an image. Specifically, we first generate a set of bounding boxes and their category labels representing spatial-semantic constraints with YOLOV3, and then obtain visual content of each bounding box with deep features extracted from a convolutional neural network. After that, we customize a similarity computation method that evaluates the relevance between dataset images and input queries according to the developed image representations. Experimental results on two large-scale benchmark retrieval datasets with images consisting of multiple objects demonstrate that our method provides an effective way to query image databases. Our code is available at https://github.com/MaJinWakeUp/spatial-content.

****

Cross-Time and Orientation-Invariant Overhead Image Geolocalization Using Deep Local Features

Yuxin Tian, Xueqing Deng, Yi Zhu, Shawn Newsam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2512-2520

Overhead image geolocalization is becoming increasingly important due to the growing collection of drone imagery without location information. In this paper, we perform large-scale overhead image geolocalization by matching a query image to wide-area reference imagery with known location. We use deep local features so that the query image need not align with but only overlap the tiled reference imagery. We further address two key challenges. For when the query and reference i

magery are from different dates, we perform cross-time geolocalization using time invariant features learned using a Siamese network. For when the query and reference imagery are oriented differently, we introduce an orientation normalization network. We demonstrate our contributions on two new high-resolution overhead image datasets. Our method significantly outperforms strong baselines on cross-time geolocalization and is shown to exhibit promising orientation invariance.
****

Geometric Image Correspondence Verification by Dense Pixel Matching
Zakaria Laskar, Iaroslav Melekhov, Hamed Rezazadegan Tavakoli, Juha Ylioinas, Juho Kannala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2521-2530

This paper addresses the problem of determining dense pixel correspondences between two images and its application to geometric correspondence verification in image retrieval. The main contribution is a geometric correspondence verification approach for re-ranking a shortlist of retrieved database images based on their dense pair-wise matching with the query image at a pixel level. We determine a set of cyclically consistent dense pixel matches between the pair of images and evaluate local similarity of matched pixels using neural network based image descriptors. Final re-ranking is based on a novel similarity function, which fuses the local similarity metric with a global similarity metric and a geometric consistency measure computed for the matched pixels. For dense matching our approach utilizes a modified version of a recently proposed dense geometric correspondence network (DGC-Net), which we also improve by optimizing the architecture. The proposed model and similarity metric compare favourably to the state-of-the-art image retrieval methods. In addition, we apply our method to the problem of long-term visual localization demonstrating promising results and generalization across datasets.
****

Image Hashing via Linear Discriminant Learning
Weixiang Hong, Yu-Ting Chang, Haifang Qin, Wei-Chih Hung, Yi-Hsuan Tsai, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2531-2539

Hashing has attracted attention in recent years due to the rapid growth of image and video data on the web. Benefiting from recent advances in deep learning, deep supervised hashing has achieved promising results for image retrieval. However, existing methods are either less efficient in data usage or incapable of learning linearly discriminative binary codes. In this paper, we revisit linear discriminative analysis and propose a linear discriminative hashing (LDH) objective that is efficient in training and achieves better accuracy in retrieval. With the joint supervision of a classification loss, we design a robust deep network to obtain binary codes that are inter-class separable and intra-class compact, which provides better representations for image retrieval. We conduct extensive experiments on three benchmark datasets, and our LDH algorithm performs favorably against existing state-of-the-art deep supervised hashing methods.
****

Stacked Adversarial Network for Zero-Shot Sketch based Image Retrieval
Anubha Pandey, Ashish Mishra, Vinay Kumar Verma, Anurag Mittal, Hema Murthy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2540-2549

Conventional approaches to Sketch-Based Image Retrieval (SBIR) assume that the data of all the classes are available during training. The assumption may not always be practical since the data of a few classes may be unavailable, or the classes may not appear at the time of training. Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR) relaxes this constraint and allows the algorithm to handle previously unseen classes during the test. This paper proposes a generative approach based on the Stacked Adversarial Network (SAN) and the advantage of Siamese Network (SN) for ZS-SBIR. While SAN generates a high-quality sample, SN learns a better distance metric compared to that of the nearest neighbor search. The capability of the generative model to synthesize image features based on the sketch reduces the SBIR problem to that of an image-to-image retrieval problem. We evaluate t he efficacy of our proposed approach on TU-Berlin, and Sketchy database in both standard ZSL and generalized ZSL setting. The proposed method yields a significant improvement in standard ZSL as well as in a more challenging generalized ZSL setting (GZSL) for SBIR.

****

2-MAP: Aligned Visualizations for Comparison of High-Dimensional Point Sets

Xiaotong Liu, Zeyu Zhang, Hong Xuan, Roxana Leontie, Abby Stylianou, Robert Pless; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2550-2558

Visualization tools like t-SNE and UMAP give insight into the high-dimensional structure of datasets. When there are related datasets (such as the high-dimensional representations of image data created by two different Deep Learning architectures), roughly aligning those visualizations helps to highlight both the similarities and differences. In this paper we propose a method to align multiple low dimensional visualizations by adding an alignment term to the UMAP loss function. We provide an automated procedure to find a weight for this term that encourages the alignment but only minimally changes the fidelity of the underlying embedding.

****

Color Composition Similarity and Its Application in Fine-grained Similarity

Mai Lan Ha, Vlad Hosu, Volker Blanz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2559-2568

Assessing visual similarity in-the-wild, a core ability of the human visual system, is a challenging problem for computer vision methods because of its subjective nature and limited annotated datasets. We make a stride forward, showing that visual similarity can be better studied by isolating its components. We identify color composition similarity as an important aspect and study its interaction with category-level similarity. Color composition similarity considers the distribution of colors and their layout in images. We create predictive models accounting for the global similarity that is beyond pixel-based and patch-based, or histogram level information. Using an active learning approach, we build a large-scale color composition similarity dataset with subjective ratings via crowd-sourcing, the first of its kind. We train a Siamese network using the dataset to create a color similarity metric and descriptors which outperform existing color descriptors. We also provide a benchmark for global color descriptors for perceptual color similarity. Finally, we combine color similarity and category level features for fine-grained visual similarity. Our proposed model surpasses the state-of-the-art performance while using three orders of magnitude less training data. The results suggest that our proposal to study visual similarity by isolating its components, modeling and combining them is a promising paradigm for further development.

****

Street Scene: A new dataset and evaluation protocol for video anomaly detection

Bharathkumar Ramachandra, Michael Jones; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2569-2578

Progress in video anomaly detection research is currently slowed by small datasets that lack a wide variety of activities as well as flawed evaluation criteria. This paper aims to help move this research effort forward by introducing a large and varied new dataset called Street Scene, as well as two new evaluation criteria that provide a better estimate of how an algorithm will perform in practice. In addition to the new dataset and evaluation criteria, we present two variations of a novel baseline video anomaly detection algorithm and show they are much more accurate on Street Scene than two well known algorithms from the literature.

****

Estimate 3D Camera Pose from 2D Pedestrian Trajectories

Yan Xu, Vivek Roy, Kris Kitani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2579-2588

We consider the task of re-calibrating the 3D pose of a static surveillance camera, whose pose may change due to external forces, such as birds, wind, falling o bjects or earthquakes. Conventionally, camera pose estimation can be solved with a PnP (Perspective-n-Point) method using 2D-to-3D feature correspondences, when 3D points are known. However, 3D point annotations are not always available or practical to obtain in real-world applications. We propose an alternative strategy for extracting 3D information to solve for camera pose by using pedestrian trajectories. We observe that 2D pedestrian trajectories indirectly contain useful 3D information that can be used for inferring camera pose. To leverage this information, we propose a data-driven approach by training a neural network (NN) regressor to model a direct mapping from 2D pedestrian trajectories projected on the image plane to 3D camera pose. We demonstrate that our regressor trained only on synthetic data can be directly applied to real data, thus eliminating the need to label any real data. We evaluate our method across six different scenes from the Town Centre Street and DUKEMTMC datasets. Our method achieves an improvement of around 50% on both position and orientation prediction accuracy when compared to other SOTA methods.

****

Detecting Face2Face Facial Reenactment in Videos

Prabhat Kumar, Mayank Vatsa, Richa Singh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2589-2597

Visual content has become the primary source of information, as evident in the billions of images and videos, shared and uploaded every single day. This has led to an increase in alterations in images and videos to make them more informative and eye-catching for the viewers worldwide. Some of these alterations are simple, like copy-move, and are easily detectable, while other sophisticated alterations like reenactment are hard to detect. Reenactment alterations allow the source to change the target expressions and create photo-realistic images and videos where such modifications are hard to detect. Significant work has been done towards creating such images and videos. However, the detection of such alterations still requires research. This research proposes a learning-based algorithm for detecting reenactment based alterations. The proposed algorithm uses a multi-stream network that learns regional artifacts and provides a robust performance at various compression levels. We also propose a loss function for the balanced learning of the streams for the proposed network. The performance is evaluated on the publicly available FaceForensics dataset. The results show state-of-the-art classification accuracy of 99.96%, 99.10%, and 91.20% for no, easy, and hard compression factors, respectively.

****

Learning a distance function with a Siamese network to localize anomalies in videos

Bharathkumar Ramachandra, Michael Jones, Ranga Vatsavai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2598-2607

This work introduces a new approach to localize anomalies in surveillance video. The main novelty is the idea of using a Siamese convolutional neural network (CNN) to learn a distance function between a pair of video patches (spatio-temporal regions of video). The learned distance function, which is not specific to the target video, is used to measure the distance between each video patch in the testing video and the video patches found in normal training video. If a testing video patch is not similar to any normal video patch then it must be anomalous. We compare our approach to previously published algorithms using 4 evaluation measures and 3 challenging target benchmark datasets. Experiments show that our approach either surpasses or performs comparably to current state-of-the-art methods.

****

Temporal Similarity Analysis of Remote Photoplethysmography for Fast 3D Mask Face Presentation Attack Detection

Siqi Liu, Xiangyuan Lan, PongChi Yuen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2608-2616

To tackle the 3D mask face presentation attack, remote Photoplethysmography (rPPG), a biomedical technique that can detect heartbeat signal remotely, is employe d as an intrinsic liveness cue. Although existing rPPG-based methods exhibit encouraging results, they require long observation time (10-12 seconds) to identify the heartbeat information, which limits their employment in real applications such as smartphone unlock and e-payment. To shorten the observation time (within 1-second) while keeping the performance, we propose a fast rPPG-based 3D mask presentation attack detection (PAD) method by analyzing the similarity of local facial rPPG signals in the time domain. In particular, a set of temporal similarity features of facial and background local rPPG signals are designed and fused to adapt the real world variations based on rPPG shape and phase properties. For better evaluation under practical variations, we build the HKBU-MARsV2+ dataset that includes 16 masks from 2 types and 6 lighting conditions. Finally, extensive experiments are conducted on 11092 short-term video slots from 4 datasets with a large number of real-world variations, in terms of mask type, lighting condition, camera, resolution of face region, and compression setting. Results show that the proposed TSrPPG outperforms the state-of-the-art competitors dramatically on discriminability and generalizability. To our best knowledge, this is the first work that addresses the length of observation time issue of rPPG-based 3D mask PAD.

****

Text-based Person Search via Attribute-aided Matching
Surbhi Aggarwal, Venkatesh Babu RADHAKRISHNAN, Anirban Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2617-2625

Text-based person search aims to retrieve the pedestrian images that best match a given text query. Existing methods utilize class-id information to get discriminative and identity-preserving features. However, it is not well-explored whether it is beneficial to explicitly ensure that the semantics of the data are retained. In the proposed work, we aim to create semantics-preserving embeddings through an additional task of attribute prediction. Since attribute annotation is typically unavailable in text-based person search, we first mine them from the text corpus. These attributes are then used as a means to bridge the modality gap between the image-text inputs, as well as to improve the representation learning. In summary, we propose an approach for text-based person search by learning an attribute-driven space along with a class-information driven space, and utilize both for obtaining the retrieval results. Our experiments on benchmark dataset, CUHK-PEDES, show that learning the attribute-space not only helps in improving performance, giving us state-of-the-art Rank-1 accuracy of 56.68%, but also yields humanly-interpretable features.

****

Multi-timescale Trajectory Prediction for Abnormal Human Activity Detection
Royston Rodrigues, Neha Bhargava, Rajbabu Velmurugan, Subhasis Chaudhuri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2626-2634

A classical approach to abnormal activity detection is to learn a representation for normal activities from the training data and then use this learned representation to detect abnormal activities while testing. Typically, the methods based on this approach operate at a fixed timescale - either a single time-instant (eg frame-based) or a constant time duration (eg video-clip based). But human abnormal activities can take place at different timescales. For example, jumping is a short-term anomaly and loitering is a long-term anomaly in a surveillance scenario. A single and pre-defined timescale is not enough to capture the wide range of anomalies occurring with different time duration. In this paper, we propose a multi-timescale model to capture the temporal dynamics at different timescales. In particular, the proposed model makes future and past predictions at different timescales for a given input pose trajectory. The model is multi-layered where intermediate layers are responsible to generate predictions corresponding to different timescales. These predictions are combined to detect abnormal activities. In addition, we also introduce a single-camera abnormal activity dataset for research use that contains 483,566 annotated frames. Our experiments show that the proposed model can capture the anomalies of different time duration and outpe rforms existing methods.

****

Relativistic Discriminator: A One-Class Classifier for Generalized Iris Presentation Attack Detection

Shivangi Yadav, Cunjian Chen, Arun Ross; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2635-2644

Iris based recognition systems are vulnerable to presentation attacks (PAs) where artifacts such as cosmetic contact lenses, artificial eyes and printed eyes can be used to fool the system. While many learning-based algorithms have been proposed to detect such attacks, very few are equipped to handle previously unseen or newly constructed PAs. In this research, we propose a presentation attack detection (PAD) method that utilizes a discriminator that is trained to distinguish between bonafide iris images and synthetically generated iris images. We hypothesize that such a discriminator will generate a tight boundary around the bonafide samples. This would allow the discriminator to better separate the bonafide samples from all types of PA samples. For generating synthetic irides, we train the Relativistic Average Standard Generative Adversarial Network (RaSGAN) that has been shown to generate higher resolution and better quality images than standard GANs. The relativistic discriminator (RD) component of the trained RaSGAN is then appropriated for PA detection and is referred to as RD-PAD. Experimental results convey the efficacy of the RD-PAD as a one-class anomaly detector.

****

Unsupervised Domain Adaptation in Person re-ID via k-Reciprocal Clustering and Large-Scale Heterogeneous Environment Synthesis

Devinder Kumar, Parthipan Siva, Paul Marchwica, Alexander Wong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2645-2654

An ongoing major challenge in computer vision is the task of person re-identification, where the goal is to match individuals across different, non-overlapping camera views. While recent success has been achieved via supervised learning using deep neural networks, such methods have limited widespread adoption due to the need for large-scale, customized data annotation. As such, there has been a re- cent focus on unsupervised learning approaches to mitigate the data annotation issue; however, current approaches in literature have limited performance compared to supervised learning approaches as well as limited applicability for adoption in new environments. In this paper, we address the aforementioned challenges faced in person re-identification for real-world, practical scenarios by introducing a novel, unsupervised domain adaptation approach for person re- identification. This is accomplished through the introduction of: i) k-reciprocal tracklet Clustering for Unsupervised Domain Adaptation (ktCUDA) (for pseudo-label generation on target domain), and ii) Synthesized Heterogeneous RE-id Domain (SHRED) composed of large-scale heterogeneous independent source environments (for improving robustness and adaptability to a wide diversity of target environments). Experimental results across four different image and video benchmark datasets show that the pro- posed ktCUDA and SHRED approach achieves an average improvement of +5.7 mAP in re-identification performance when compared to existing state-of-the-art methods, as well as demonstrate better adaptability to different types of environments.

****

Video Person Re-Identification using Learned Clip Similarity Aggregation

Neeraj Matiyali, Gaurav Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2655-2664

We address the challenging task of video-based person re-identification. Recent works have shown that splitting the video sequences into clips and then aggregating clip based similarity is appropriate for the task. We show that using a learned clip similarity aggregation function allows filtering out hard clip pairs, e.g. where the person is not clearly visible, is in a challenging pose, or where the poses in the two clips are too different to be informative. This allows the method to focus on clip-pairs which are more informative for the task. We also introduce the use of 3D CNNs for video-based re-identification and show their eff ectiveness by performing equivalent to previous works, which use optical flow in addition to RGB, while using RGB inputs only. We give quantitative results on three challenging public benchmarks and show better or competitive performance. We also validate our method qualitatively.

****

SmoothFool: An Efficient Framework for Computing Smooth Adversarial Perturbations

Ali Dabouei, Sobhan Soleymani, Fariborz Taherkhani, Jeremy Dawson, Nasser Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2665-2674

Deep neural networks are susceptible to adversarial manipulations in the input domain. The extent of vulnerability has been explored intensively in cases of l_p-bounded and l_p-minimal adversarial perturbations. However, the vulnerability of DNNs to adversarial perturbations with specific statistical properties or frequency-domain characteristics has not been sufficiently explored. In this paper, we study the smoothness of perturbations and propose SmoothFool, a general and computationally efficient framework for computing smooth adversarial perturbations. Through extensive experiments, we validate the efficacy of the proposed method for both the white-box and black-box attack scenarios. In particular, we demonstrate that: (i) there exist extremely smooth adversarial perturbations for well-established and widely used network architectures, (ii) smoothness significantly enhances the robustness of perturbations against state-of-the-art defense mechanisms, (iii) smoothness improves the transferability of adversarial perturbations across both data points and network architectures, and (iv) class categories exhibit a variable range of susceptibility to smooth perturbations. Our results suggest that smooth APs can play a significant role in exploring the vulnerability extent of DNNs to adversarial examples.

****

Pose Guided Gated Fusion for Person Re-identification

Amran Bhuiyan, Yang Liu, Parthipan Siva, Mehrsan Javan, Ismail Ben Ayed, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2675-2684

Person re-identification is an important yet challenging problem in visual recognition. Despite the recent advances with deep learning (DL) models for spatio-temporal and multi-modal fusion, re-identification approaches often fail to leverage the contextual information (e.g., pose and illu- mination) to dynamically select the most discriminant con- volutional filters (i.e., appearance features) for feature rep- resentation and inference. State-of-the-art techniques for gated fusion employ complex dedicated part- or attention- based architectures for late fusion, and do not incorpo- rate pose and appearance information to train the back- bone network. In this paper, a new DL model is proposed for pose-guided re-identification, comprised of a deep back- bone, pose estimation, and gated fusion network. Given a query image of an individual, the backbone convolutional NN produces a feature embedding required for pair-wise matching with embeddings for reference images, where fea- ture maps from the pose network and from mid-level CNN layers are combined by the gated fusion network to gen- erate pose-guided gating. The proposed framework al- lows to dynamically activate the most discriminant CNN filters based on pose information in order to perform a finer grained recognition. Extensive experiments on three challenging benchmark datasets indicate that integrating the pose-guided gated fusion into the state-of-the-art re- identification backbone architecture allows to improve their recognition accuracy. Experimental results also support our intuition on the advantages of gating backbone appear- ance information using the pose feature maps at mid-level CNN layers.

****

EDGE20: A Cross Spectral Evaluation Dataset for Multiple Surveillance Problems

Ha Le, Christos Smailis, Lei Shi, Ioannis Kakadiaris; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2685-2694

Surveillance-related datasets that have been released in recent years focus only on one specific problem at a time (e.g., pedestrian detection, face detection, or face recognition), while most of them were collected using visible spectrum (VIS) cameras. Even though some cross-spectral datasets were presented in the past, they were acquired in a constrained setup, which limited the performance of methods for the aforementioned problems under a cross-spectral setting. This work introduces a new dataset, named EDGE20, that can be used in addressing the problems of pedestrian detection, face detection, and face recognition in images captured using trail cameras under the VIS and NIR spectra. Data acquisition was performed in an outdoor environment, during both day and night, under unconstrained acquisition conditions. The collection of images is accompanied by a rich set of annotations, consisting of person and facial bounding boxes, unique subject identifiers, and labels that characterize facial images as frontal, profile, or back faces. Moreover, the performance of several state-of-the-art methods was evaluated for each of the scenarios covered by our dataset. The baseline results we obtained highlight the difficulty of current methods in the tasks of cross-spectral pedestrian detection, face detection, and face recognition due to unconstrained conditions, including low resolution, pose variation, illumination variation, occlusions, and motion blur.

****

DeFraudNet:End2End Fingerprint Spoof Detection using Patch Level Attention

Anusha B, Sayan Banerjee, Subhasis Chaudhuri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2695-2704

In recent years, fingerprint recognition systems have made remarkable advancements in the field of biometric security as it plays an important role in personal, national and global security. In spite of all these notable advancements, the fingerprint recognition technology is still susceptible to spoof attacks which can significantly jeopardize the user security. The cross sensor and cross material spoof detection still pose a challenge with a myriad of spoof materials emerging every day, compromising sensor interoperability and robustness. This paper proposes a novel method for fingerprint spoof detection using both global and local fingerprint feature descriptors. These descriptors are extracted using DenseNet which significantly improves cross-sensor, cross-material and cross-dataset performance. A novel patch attention network is used for finding the most discriminative patches and also for network fusion. We evaluate our method on four publicly available datasets: LivDet 2011, 2013, 2015 and 2017. A set of comprehensive experiments are carried out to evaluate cross-sensor, cross-material and cross-dataset performance over these datasets. The proposed approach achieves an average accuracy of 99.52%, 99.16% and 99.72% on LivDet 2017, 2015 and 2011 respectively outperforming the current state-of-the-art results by 3% and 4% for LivDet 2015 and 2011 respectively.

****

Devon: Deformable Volume Network for Learning Optical Flow

Yao Lu, Jack Valmadre, Heng Wang, Juho Kannala, Mehrtash Harandi, Philip Torr; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2705-2713

State-of-the-art neural network models estimate large displacement optical flow in multi-resolution and use warping to propagate the estimation between two resolutions. Despite their impressive results, it is known that there are two problems with the approach. First, the multi-resolution estimation of optical flow fails in situations where small objects move fast. Second, warping creates artifacts when occlusion or dis-occlusion happens. In this paper, we propose a new neural network module, Deformable Cost Volume, which alleviates the two problems. Based on this module, we designed the Deformable Volume Network (Devon) which can estimate multi-scale optical flow in a single high resolution. Experiments show Devon is more suitable in handling small objects moving fast and achieves comparable results to the state-of-the-art methods in public benchmarks.

****

Stochastic Dynamics for Video Infilling

Qiangeng Xu, Hanwang Zhang, Weiyue Wang, Peter Belhumeur, Ulrich Neumann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (

WACV), 2020, pp. 2714-2723

In this paper, we introduce a stochastic dynamics video infilling (SDVI) framework to generate frames between long intervals in a video. Our task differs from video interpolation which aims to produce transitional frames for a short interval between every two frames and increase the temporal resolution. Our task, namely video infilling, however, aims to infill long intervals with plausible frame sequences. Our framework models the infilling as a constrained stochastic generation process and sequentially samples dynamics from the inferred distribution. SDVI consists of two parts: (1) a bi-directional constraint propagation module to guarantee the spatial-temporal coherence among frames, (2) a stochastic sampling process to generate dynamics from the inferred distributions. Experimental results show that SDVI can generate clear frame sequences with varying contents. Moreover, motions in the generated sequence are realistic and able to transfer smoothly from the given start frame to the terminal frame.

****

Cross-Conditioned Recurrent Networks for Long-Term Synthesis of Inter-Person Human Motion Interactions

Jogendra Nath Kundu, Himanshu Buckchash, Priyanka Mandikal, Rahul M V, Anirudh Jamkhandi, Venkatesh Babu RADHAKRISHNAN; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2724-2733

Modeling dynamics of human motion is one of the most challenging sequence modeling problem, with diverse applications in animation industry, human-robot interaction, motion-based surveillance, etc. Available attempts to use auto-regressive techniques for long-term single-person motion generation usually fails, resulting in stagnated motion or divergence to unrealistic pose patterns. In this paper, we propose a novel cross-conditioned recurrent framework targeting long-term synthesis of inter-person interactions beyond several minutes. We carefully integrate positive implications of both auto-regressive and encoder-decoder recurrent architecture, by interchangeably utilizing two separate fixed-length cross person motion prediction models for long-term generation in a novel hierarchical fashion. As opposed to prior approaches, we guarantee structural plausibility of 3D pose by training the recurrent model to regress latent representation of a separately trained generative pose embedding network. Different variants of the proposed frameworks are evaluated through extensive experiments on SBU-interaction, CMU-MoCAP and an inhouse collection of duet-dance dataset. Qualitative and quantitative evaluation on several tasks, such as Short-term motion prediction, Long-term motion synthesis and Interaction-based motion retrieval against prior state-of-the-art approaches clearly highlight superiority of the proposed framework.

****

MotionRec: A Unified Deep Framework for Moving Object Recognition

Murari Mandal, Lav Kush Kumar, Mahipal Singh Saran, Santosh Kumar vipparthi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2734-2743

In this paper we present a novel deep learning framework to perform online moving object recognition (MOR) in streaming videos. The existing methods for moving object detection (MOD) only computes class-agnostic pixel-wise binary segmentation of video frames. On the other hand, the object detection techniques do not differentiate between static and moving objects. To the best of our knowledge, this is a first attempt for simultaneous localization and classification of moving objects in a video, i.e. MOR in a single-stage deep learning framework. We achieve this by labelling axis-aligned bounding boxes for moving objects which requires less computational resources than producing pixel-level estimates. In the proposed MotionRec, both temporal and spatial features are learned using past history and current frames respectively. First, the background is estimated with a temporal depth reductionist (TDR) block. Then the estimated background, current frame and temporal median of recent observations are assimilated to encode spatiotemporal motion saliency. Moreover, feature pyramids are generated from these motion saliency maps to perform regression and classification at multiple levels of feature abstractions. MotionRec works online at inference as it requires only few past frames for MOR. Moreover, it doesn't require predefined target initializ ation from user. We also annotated axis-aligned bounding boxes (42,614 objects (14,814 cars and 27,800 person) in 24,923 video frames of CDnet 2014 dataset) due to lack of available benchmark datasets for MOR. The performance is observed qualitatively and quantitatively in terms of mAP over a defined unseen test set. Experiments show that the proposed MotionRec significantly improves over strong baselines with RetinaNet architectures for MOR.

****

TwoStreamVAN: Improving Motion Modeling in Video Generation

Ximeng Sun, Huijuan Xu, Kate Saenko; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2744-2753

Video generation is an inherently challenging task, as it requires modeling realistic temporal dynamics as well as spatial content. Existing methods entangle the two intrinsically different tasks of motion and content creation in a single generator network, but this approach struggles to simultaneously generate plausible motion and content. To im-prove motion modeling in video generation tasks, we propose a two-stream model that disentangles motion generation from content generation, called a Two-Stream Variational Adversarial Network (TwoStreamVAN). Given an action label and a noise vector, our model is able to create clear and consistent motion, and thus yields photorealistic videos. The key idea is to progressively generate and fuse multi-scale motion with its corresponding spatial content. Our model significantly outperforms existing methods on the standard Weizmann Human Action, MUG Facial Expression, and VoxCeleb datasets, as well as our new dataset of diverse human actions with challenging and complex motion. Our code is available at https://github.com/sunxm2357/TwoStreamVAN/.

****

NRMVS: Non-Rigid Multi-view Stereo

Matthias Innmann, Kihwan Kim, Jinwei Gu, Matthias Niessner, Charles Loop, Marc Stamminger, Jan Kautz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2754-2763

Multi-view Stereo (MVS) is a common solution in photogrammetry applications for the dense reconstruction of a static scene from images. The static scene assumption, however, limits the general applicability of MVS algorithms, as many day-to-day scenes undergo non-rigid motion, e.g., clothes, faces, or human bodies. In this paper, we open up a new challenging direction: Dense 3D reconstruction of scenes with non-rigid changes observed from a small number of images sparsely captured from different views with a single monocular camera, which we call non-rigid multi-view stereo (NRMVS) problem. We formulate this problem as a joint optimization of deformation and depth estimation, using deformation graphs as the underlying representation. We propose a new sparse 3D to 2D matching technique with a dense patch-match evaluation scheme to estimate the most plausible deformation field satisfying depth and photometric consistency. We show that a dense reconstruction of a scene with non-rigid changes from a few images is possible, and demonstrate that our method can be used to interpolate novel deformed scenes from various combinations of deformation estimates derived from the sparse views.

****

Fusing Semantics and Motion State Detection for Robust Visual SLAM

Gaurav Singh, Meiqing Wu, Siew-Kei Lam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2764-2773

Achieving robust pose tracking and mapping in highly dynamic environments is a major challenge faced by existing visual SLAM (vSLAM) systems. In this paper, we increase the robustness of existing vSLAM by accurately removing moving objects from the scene so that they will not contribute to pose estimation and mapping. Specifically, semantic information is fused with motion states of the scene via a probability framework to enable accurate and robust moving object extraction in order to retain the useful features for pose estimation and mapping. Our work highlights the importance of distinguishing between motion states of potential moving objects for vSLAM in highly dynamic environments. The proposed method can be integrated into existing vSLAM systems to increase their robustness in dynamic environments without incurring much computation cost. We provide extensive experimental results on three well-known datasets to show that the proposed techniq ue outperforms existing vSLAM methods in indoor and outdoor environments, under various scenarios such as crowded scenes.

****

BSUV-Net: A Fully-Convolutional Neural Network for Background Subtraction of Unseen Videos

Ozan Tezcan, Prakash Ishwar, Janusz Konrad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2774-2783

Background subtraction is a basic task in computer vision and video processing often applied as a pre-processing step for object tracking, people recognition, etc. Recently, a number of successful background-subtraction algorithms have been proposed, however nearly all of the top-performing ones are supervised. Crucially, their success relies upon the availability of some annotated frames of the test video during training. Consequently, their performance on completely "unseen" videos is undocumented in the literature. In this work, we propose a new, supervised, background-subtraction algorithm for unseen videos (BSUV-Net) based on a fully-convolutional neural network. The input to our network consists of the current frame and two background frames captured at different time scales along with their semantic segmentation maps. In order to reduce the chance of overfitting, we also introduce a new data-augmentation technique which mitigates the impact of illumination difference between the background frames and the current frame. On the CDNet-2014 dataset, BSUV-Net outperforms state-of-the-art algorithms evaluated on unseen videos in terms of several metrics including F-measure, recall and precision.

****

Disentangling Human Dynamics for Pedestrian Locomotion Forecasting with Noisy Supervision

Karttikeya Mangalam, Ehsan Adeli, Kuan-Hui Lee, Adrien Gaidon, Juan Carlos Niebles; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2784-2793

We tackle the problem of Human Locomotion Forecasting, a task for jointly predicting the spatial positions of several keypoints on human body in the near future under an egocentric setting. In contrast to the previous work that aims to solve either the task of pose prediction or trajectory forecasting in isolation, we propose a framework to unify these two problems and address the practically useful task of pedestrian locomotion prediction in the wild. Among the major challenges in solving this task is the scarcity of annotated egocentric video datasets with dense annotations for pose, depth, or egomotion.   To surmount this difficulty, we use state-of-the-art models to generate (noisy) annotations and propose robust forecasting models that can learn from this noisy supervision. We present a method to disentangle the overall pedestrian motion into easier to learn subparts by utilizing a pose completion and a decomposition module. The completion module fills in the missing key-point annotations and the decomposition module breaks the cleaned locomotion down to global (trajectory) and local (pose keypoint movements). Further, with Quasi RNN as our backbone, we propose a novel hierarchical trajectory forecasting network that utilizes low-level vision domain specific signals like egomotion and depth to predict the global trajectory. Our method leads to state-of-the-art results for the prediction of human locomotion in the egocentric view

****

Adapting Grad-CAM for Embedding Networks

Lei Chen, Jianhui Chen, Hossein Hajimirsadeghi, Greg Mori; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2794-2803

The gradient-weighted class activation mapping (Grad-CAM) method can faithfully highlight important regions in images for deep model prediction in image classification, image captioning and many other tasks. It uses the gradients in back-propagation as weights (grad-weights) to explain network decisions. However, applying Grad-CAM to embedding networks raises significant challenges because embedding networks are trained by millions of dynamically paired examples (e.g. triplets). To overcome these challenges, we propose an adaptation of the Grad-CAM metho

d for embedding networks. First, we aggregate grad-weights from multiple training examples to improve the stability of Grad-CAM. Then, we develop an efficient weight-transfer method to explain decisions for any image without back-propagation. We extensively validate the method on the standard CUB200 dataset in which our method produces more accurate visual attention than the original Grad-CAM method. We also apply the method to a house price estimation application using images. The method produces convincing qualitative results, showcasing the practicality of our approach.
****
Super-resolved Chromatic Mapping of Snapshot Mosaic Image Sensors via a Texture Sensitive Residual Network
Mehrdad Shoeiby, Lars Petersson, Ali Armin, Sadegh Aliakbarian, antonio robbles-kelly; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2804-2813
This paper introduces a novel method to simultaneously super-resolve and colour-predict images acquired by snapshot mosaic sensors. These sensors allow for spectral images to be acquired using low-power, small form factor, solid-state CMOS sensors that can operate at video frame rates without the need for complex optical setups. Despite their desirable traits, their main drawback stems from the fact that the spatial resolution of the imagery acquired by these sensors is low. Moreover, chromatic mapping in snapshot mosaic sensors is not straightforward since the bands delivered by the sensor tend to be narrow and unevenly distributed across the range in which they operate. We tackle this drawback as applied to chromatic mapping by using a residual channel attention network equipped with a texture sensitive block. Our method significantly outperforms the traditional approach of interpolating the image and, afterwards, applying a colour matching function. This work establishes state-of-the-art in this domain while also making available to the research community a dataset containing 296 registered stereo multi-spectral/RGB images pairs.
****
Few-Shot Scene Adaptive Crowd Counting Using Meta-Learning
Mahesh Kumar Krishna Reddy, Mohammad Hossain, Mrigank Rochan, Yang Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2814-2823
We consider the problem of few-shot scene adaptive crowd counting. Given a target camera scene, our goal is to adapt a model to this specific scene with only a few labeled images of that scene. The solution to this problem has potential applications in numerous real-world scenarios, where we ideally like to deploy a crowd counting model specially adapted to a target camera. We accomplish this challenge by taking inspiration from the recently introduced learning-to-learn paradigm in the context of few-shot regime. In training, our method learns the model parameters in a way that facilitates the fast adaptation to the target scene. At test time, given a target scene with a small number of labeled data, our method quickly adapts to that scene with a few gradient updates to the learned parameters. Our extensive experimental results show that the proposed approach outperforms other alternatives in few-shot scene adaptive crowd counting.
****
PointPoseNet: Point Pose Network for Robust 6D Object Pose Estimation
Wei Chen, Jinming Duan, Hector Basevi, Hyung Jin Chang, Ales Leonardis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2824-2833
In this paper, we propose a novel pipeline to estimate 6D object pose from RGB-D images of known objects present in complex scenes. The pipeline directly operates on raw point clouds extracted from RGB-D scans. Specifically, our method takes the point cloud as input and regresses the point-wise unit vectors pointing to the 3D keypoints. We then use these vectors to generate keypoint hypotheses from which the 6D object pose hypotheses are computed. Finally, we select the best 6D object pose from the hypotheses based on a proposed scoring mechanism with geometry constraints. Extensive experiments show that the proposed method is robust against the variety in object shape and appearance as well as occlusions betwe

en objects, and that our method outperforms the state-of-the-art methods on the LINEMOD and Occlusion LINEMOD datasets.

****

Predicting the Physical Dynamics of Unseen 3D Objects
Davis Rempe, Srinath Sridhar, He Wang, Leonidas Guibas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2834-2843

Machines that can predict the effect of physical interactions on the dynamics of previously unseen object instances are important for creating better robots and interactive virtual worlds. In this work, we focus on predicting the dynamics of 3D objects on a plane that have just been subjected to an impulsive force. In particular, we predict the changes in state - 3D position, rotation, velocities, and stability. Different from previous work, our approach can generalize dynamics predictions to object shapes and initial conditions that were unseen during training. Our method takes the 3D object's shape as a point cloud and its initial linear and angular velocities as input. We extract shape features and use a recurrent neural network to predict the full change in state at each time step. Our model can support training with data from both a physics engine or the real world. Experiments show that we can accurately predict the changes in state for unseen object geometries and initial conditions.

****

Distributed Iterative Gating Networks for Semantic Segmentation
Rezaul Karim, Md Amirul Islam, Neil D. B. Bruce; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2844-2853

In this paper, we present a canonical structure for controlling information flow in neural networks with an efficient feedback routing mechanism based on a strategy of Distributed Iterative Gating (DIGNet). The structure of this mechanism derives from a strong conceptual foundation, and presents a light-weight mechanism for adaptive control of computation similar to recurrent convolutional neural networks by integrating feedback signals with a feed forward architecture. In contrast to other RNN formulations, DIGNet generates feedback signals in a cascaded manner that implicitly carries information from all the layers above. This cascaded feedback propagation by means of the propagator gates is found to be more effective compared to other feedback mechanisms that use feedback from output of either the corresponding stage or from the previous stage. Experiments reveal the high degree of capability that this recurrent approach with cascaded feedback presents over feed-forward baselines and other recurrent models for pixel-wise labeling problems on three challenging datasets, PASCAL VOC 2012, COCO-Stuff, and ADE20K.

****

Audio-Visual Model Distillation Using Acoustic Images
Andres Perez, Valentina Sanguineti, Pietro Morerio, Vittorio Murino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2854-2863

In this paper, we investigate how to learn rich and robust feature representations for audio classification from visual data and acoustic images, a novel audio data modality. Former models learn audio representations from raw signals or spectral data acquired by a single microphone, with remarkable results in classification and retrieval. However, such representations are not so robust towards variable environmental sound conditions. We tackle this drawback by exploiting a new multimodal labeled action recognition dataset acquired by a hybrid audio-visual sensor that provides RGB video, raw audio signals, and spatialized acoustic data, also known as acoustic images, where the visual and acoustic images are aligned in space and synchronized in time. Using this richer information, we train audio deep learning models in a teacher-student fashion. In particular, we distill knowledge into audio networks from both visual and acoustic image teachers. Our experiments suggest that the learned representations are more powerful and have better generalization capabilities than the features learned from models trained using just single-microphone audio data.

****

Going Beyond the Regression Paradigm with Accurate Dot Prediction for Dense Crowds

deepak babu sam, Skand Peri, Mukuntha Narayanan Sundararaman, Venkatesh Babu RADHAKRISHNAN; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2864-2872

We present an alternative to the paradigm of density regression widely being employed for tackling crowd counting. In the prevalent regression approach, a model is trained for mapping images to its crowd density rather than counting by detecting every person. This framework is motivated from the difficulty to discriminate humans in highly dense crowds where unfavorable perspective, occlusion and clutter are prevalent. Though regression methods estimate overall crowd counts pretty well, localization of individual persons suffers and varies considerably across the entire density spectrum. Moreover, individual detection of people aids more explainable practical systems than predicting blind crowd count or density map. Hence, we move away from density regression and reformulate the task as localized dot prediction in dense crowds. Our dot detection model, DD-CNN, is trained for pixel-wise binary classification to detect people instead of regressing local crowd density. In order to handle severe scale variation and detect people of all scales with accurate dots, we use a novel multi-scale architecture which does not require any ground truth scale information. This training regime, which incorporates top-down feedback, helps our model to localize people in sparse as well as dense crowds. Our model delivers superior counting performance on major crowd datasets. We also evaluate on some additional metrics and evidence superior localization of the dot detection formulation.

****

Efficient Video Semantic Segmentation with Labels Propagation and Refinement

Matthieu Paul, Christoph Mayer, Luc Van Gool, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2873-2882

This paper tackles the problem of real-time semantic segmentation of high definition videos using a hybrid GPU / CPU approach. We propose an Efficient Video Segmentation (EVS) pipeline that combines: (i) On the CPU, a very fast optical flow method, that is used to exploit the temporal aspect of the video and propagate semantic information from one frame to the next. It runs in parallel with the GPU. (ii) On the GPU, two Convolutional Neural Networks: A main segmentation network that is used to predict dense semantic labels from scratch, and a Refiner that is designed to improve predictions from previous frames with the help of a fast Inconsistencies Attention Module (IAM). The latter can identify regions that cannot be propagated accurately. We suggest several operating points depending on the desired frame rate and accuracy. Our pipeline achieves accuracy levels competitive to the existing real-time methods for semantic image segmentation(mIoU above 60%), while achieving much higher frame rates. On the popular Cityscapes dataset with high resolution frames (2048 x 1024), the proposed operating points range from 80 to 1000 Hz on a single GPU and CPU.

****

Plugin Networks for Inference under Partial Evidence

Michal Koperski, Tomasz Konopczynski, Rafal Nowak, Piotr Semberecki, Tomasz Trzcinski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2883-2891

In this paper, we propose a novel method to incorporate partial evidence in the inference of deep convolutional neural networks. Contrary to the existing, top-performing methods, which either iteratively modify the input of the network or exploit external label taxonomy to take the partial evidence into account, we add separate network modules ("Plugin Networks") to the intermediate layers of a pre-trained convolutional network. The goal of these modules is to incorporate additional signal, i.e. information about known labels, into the inference procedure, and adjust the predicted output accordingly. Since the attached plugins have a simple structure, consisting of only fully connected layers, we drastically reduced the computational cost of training and inference. Also, the proposed architecture allows propagating information about known labels directly to the interm ediate layers to improve the final representation. Extensive evaluation of the proposed method confirms that our Plugin Networks outperform the state-of-the-art in a variety of tasks, including scene categorization, multi-label image annotation, and semantic segmentation.

****

Classifying All Interacting Pairs in a Single Shot
Sanaa Chafik, Astrid Orcesi, Romaric Audigier, Bertrand Luvison; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2892-2901

In this paper, we introduce a novel human interaction detection approach, based on CALIPSO (Classifying ALl Interacting Pairs in a Single shOt), a classifier of human-object interactions. This new single-shot interaction classifier estimates interactions simultaneously for all human-object pairs, regardless of their number and class. State-of-the-art approaches adopt a multi-shot strategy based on a pairwise estimate of interactions for a set of human-object candidate pairs, which leads to a complexity depending, at least, on the number of interactions or, at most, on the number of candidate pairs. In contrast, the proposed method estimates the interactions on the whole image. Indeed, it simultaneously estimates all interactions between all human subjects and object targets by performing a single forward pass throughout the image. Consequently, it leads to a constant complexity and computation time independent of the number of subjects, objects or interactions in the image. In detail, interaction classification is achieved on a dense grid of anchors thanks to a joint multi-task network that learns three complementary tasks simultaneously: (i) prediction of the types of interaction, (ii) estimation of the presence of a target and (iii) learning of an embedding which maps interacting subject and target to a same representation, by using a metric learning strategy. In addition, we introduce an object-centric passive-voice verb estimation which significantly improves results. Evaluations on the two well-known Human-Object Interaction image datasets, V-COCO and HICO-DET, demonstrate the competitiveness of the proposed method (2nd place) compared to the state-of-the-art while having constant computation time regardless of the number of objects and interactions in the image.

****

A Little Fog for a Large Turn
Harshitha Machiraju, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2902-2911

Small, carefully crafted perturbations called adversarial perturbations can easily fool neural networks. However, these perturbations are largely additive and not naturally found. We turn our attention to the field of Autonomous navigation wherein adverse weather conditions such as fog have a drastic effect on the predictions of these systems. These weather conditions are capable of acting like natural adversaries that can help in testing models. To this end, we introduce a general notion of adversarial perturbations, which can be created using generative models and provide a methodology inspired by Cycle-Consistent Generative Adversarial Networks to generate adversarial weather conditions for a given image. Our formulation and results show that these images provide a suitable testbed for steering models used in Autonomous navigation models. Our work also presents a more natural and general definition of Adversarial perturbations based on Perceptual Similarity.

****

Smart Hypothesis Generation for Efficient and Robust Room Layout Estimation
Martin Hirzer, Vincent Lepetit, PETER ROTH; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2912-2920

We propose a novel method to efficiently estimate the spatial layout of a room from a single monocular RGB image. As existing approaches based on low-level feature extraction, followed by a vanishing point estimation are very slow and often unreliable in realistic scenarios, we build on semantic segmentation of the input image. To obtain better segmentations, we introduce a robust, accurate and very efficient hypothesize-and-test scheme. The key idea is to use three segmentation hypotheses, each based on a different number of visible walls. For each hypo thesis, we predict the image locations of the room corners and select the hypothesis for which the layout estimated from the room corners is consistent with the segmentation. We demonstrate the efficiency and robustness of our method on three challenging benchmark datasets, where we significantly outperform the state-of-the-art.

****

Graph Neural Networks for Image Understanding Based on Multiple Cues: Group Emotion Recognition and Event Recognition as Use Cases

Xin Guo, Luisa Polania, Bin Zhu, Charles Boncelet, Kenneth Barner; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2921-2930

A graph neural network (GNN) for image understanding based on multiple cues is proposed in this paper. Compared to traditional feature and decision fusion approaches that neglect the fact that features can interact and exchange information, the proposed GNN is able to pass information among features extracted from different models. Two image understanding tasks, namely group-level emotion recognition (GER) and event recognition, which are highly semantic and require the interaction of several deep models to synthesize multiple cues, were selected to validate the performance of the proposed method. It is shown through experiments that the proposed method achieves state-of-the-art performance on the selected image understanding tasks. In addition, a new group-level emotion recognition database is introduced and shared in this paper.

****

Global Context Reasoning for Semantic Segmentation of 3D Point Clouds

Yanni Ma, Yulan Guo, Hao Liu, Yinjie Lei, Gongjian Wen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2931-2940

Global contextual dependency is important for semantic segmentation of 3D point clouds. However, most existing approaches stack feature extraction layers to enlarge the receptive field to aggregate more contextual information of points along the spatial dimension. In this paper, we propose a Point Global Context Reasoning (PointGCR) module to capture global contextual information along the channel dimension. In PointGCR, an undirected graph representation (namely, ChannelGraph) is used to learn channel independencies. Specifically, channel maps are first represented as graph nodes and the independencies between nodes are then represented as graph edges. PointGCR is a plug-andplay and end-to-end trainable module. It can easily be integrated into an existing segmentation network and achieves a significant performance improvement. We conduct extensive experiments to evaluate the proposed PointGCR module on both indoor and outdoor datasets. Experimental results show that our PointGCR module efficiently captures global contextual dependencies and significantly improve the segmentation performance of several existing networks.

****

Combinational Class Activation Maps for Weakly Supervised Object Localization

Seunghan Yang, Yoonhyung Kim, Youngeun Kim, Changick Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2941-2949

Weakly supervised object localization has recently attracted attention since it aims to identify both class labels and locations of objects by using image-level labels. Most previous methods utilize the activation map corresponding to the highest activation source. Exploiting only one activation map of the highest probability class is often biased into limited regions or sometimes even highlights background regions. To resolve these limitations, we propose to use activation maps, named combinational class activation maps (CCAM), which are linear combinations of activation maps from the highest to the lowest probability class. By using CCAM for localization, we suppress background regions to help highlighting foreground objects more accurately. In addition, we design the network architecture to consider spatial relationships for localizing relevant object regions. Specifically, we integrate non-local modules into an existing base network at both low- and high-level layers. Our final model, named non-local combinational class

activation maps (NL-CCAM), obtains superior performance compared to previous methods on representative object localization benchmarks including ILSVRC 2016 and CUB-200-2011. Furthermore, we show that the proposed method has a great capability of generalization by visualizing other datasets.
****

Fine-grained Image Classification and Retrieval by Combining Visual and Locally Pooled Textual Features

Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluis Gomez, Dimosthenis Karatzas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2950-2959

Text contained in an image carries high-level semantics that can be exploited to achieve richer image understanding. In particular, the mere presence of text provides strong guiding content that should be employed to tackle a diversity of computer vision tasks such as image retrieval, fine-grained classification, and visual question answering. In this paper, we address the problem of fine-grained classification and image retrieval by leveraging textual information along with visual cues to comprehend the existing intrinsic relation between the two modalities. The novelty of the proposed model consists of the usage of a PHOC descriptor to construct a bag of textual words along with a Fisher Vector Encoding that captures the morphology of text. This approach provides a stronger multimodal representation for this task and as our experiments demonstrate, it achieves state-of-the-art results on two different tasks, fine-grained classification and image retrieval.
****

Iterative and Adaptive Sampling with Spatial Attention for Black-Box Model Explanations

Bhavan Vasu, Chengjiang Long; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2960-2969

Deep neural networks have achieved great success in many real-world applications, yet it remains unclear and difficult to explain their decision-making process to an end user. In this paper, we address the explainable AI problem for deep neural networks with our proposed framework, named IASSA that generates an importance map indicating how salient each pixel is for the model's prediction with an iterative and adaptive sampling module. We employ an affinity matrix calculated on multi-level deep learning features to explore long-range pixel-to-pixel correlation, which can shift the saliency values guided by our long-range and parameter-free spatial attention. Extensive experiments on the MS-COCO dataset show that our proposed approach matches or exceeds the performance of state-of-the-art black-box explanation methods.
****

See the Sound, Hear the Pixels

Janani Ramaswamy, Sukhendu Das; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2970-2979

For every event occurring in the real world, most often a sound is associated with the corresponding visual scene. Humans possess an inherent ability to automatically map the audio content with visual scenes leading to an effortless and enhanced understanding of the underlying event. This triggers an interesting question: Can this natural correspondence between video and audio, which has been diminutively explored so far, be learned by a machine and modeled jointly to localize the sound source in a visual scene? In this paper, we propose a novel algorithm that addresses the problem of localizing sound source in unconstrained videos, which uses efficient fusion and attention mechanisms. Two novel blocks namely, Audio Visual Fusion Block (AVFB) and Segment-Wise Attention Block (SWAB) have been developed for this purpose. Quantitative and qualitative evaluations show that it is feasible to use the same algorithm with minor modifications to serve the purpose of sound localization using three different types of learning: supervised, weakly supervised and unsupervised. A novel Audio Visual Triplet Gram Matrix Loss (AVTGML) has been proposed as a loss function to learn the localization in an unsupervised way. Our empirical evaluations demonstrate a significant increase in performance over the existing state-of-the-art methods, serving as a testi mony to the superiority of our proposed approach.

****

Focusing Visual Relation Detection on Relevant Relations with Prior Potentials

Francois PLESSE, Alexandru Ginsca, Bertrand DELEZOIDE, Francoise PRETEUX; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2980-2989

Understanding images relies on the understanding of how visible objects are linked to each other. Current approaches of Visual Relation Detection (VRD) are hindered by the high frequency of some relations: when an important focus is put on them, more meaningful ones are overlooked. We address this challenge by learning the relative relevance of relations, and integrating this term into a novel scene graph extraction scheme. We show that this allows our model to predict relations on fewer and more relevant object pairs. It outperforms MotifNet, a state of the art model, on the Visual Genome dataset. It increases the Class Macro recall, the metric we propose to use, from 38.1% to 44.4%. In addition, we propose a new split of Visual Genome, with a more balanced relation distribution, emphasizing on the detection of uncommon relations and validates the use of the previous metric. On this set, our model outperforms MotifNet on all metrics, e.g. from 39.6% to 44.0% at 10 predictions per image on the relation classification task.

****

Domain Bridge for Unpaired Image-to-Image Translation and Unsupervised Domain Adaptation

Fabio Pizzati, Raoul de Charette, Michela Zaccaria, Pietro Cerri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2990-2998

Image-to-image translation architectures may have limited effectiveness in some circumstances. For example, while generating rainy scenarios, they may fail to model typical traits of rain as water drops, and this ultimately impacts the synthetic images realism. With our method, called domain bridge, web-crawled data are exploited to reduce the domain gap, leading to the inclusion of previously ignored elements in the generated images. We make use of a network for clear to rain translation trained with the domain bridge to extend our work to Unsupervised Domain Adaptation (UDA). In that context, we introduce an online multimodal style-sampling strategy, where image translation multimodality is exploited at training time to improve performances. Finally, a novel approach for self-supervised learning is presented, and used to further align the domains. With our contributions, we simultaneously increase the realism of the generated images, while reaching on par performances with respect to the UDA state-of-the-art, with a simpler approach.

****

Munich to Dubai: How far is it for Semantic Segmentation?

Shyam Nandan Rai, Vineeth N Balasubramanian, Anbumani Subramanian, C.V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 2999-3008

Cities having hot weather conditions results in geometrical distortion, thereby adversely affecting the performance of semantic segmentation model. In this work, we study the problem of semantic segmentation model in adapting to such hot climate cities. This issue can be circumvented by collecting and annotating images in such weather conditions and training segmentation models on those images. But the task of semantically annotating images for every environment is painstaking and expensive. Hence, we propose a framework that improves the performance of semantic segmentation models without explicitly creating an annotated dataset for such adverse weather variations. Our framework consists of two parts, a restoration network to remove the geometrical distortions caused by hot weather and an adaptive segmentation network that is trained on an additional loss to adapt to the statistics of the ground-truth segmentation map. We train our framework on the Cityscapes dataset, which showed a total IoU gain of 12.707 over standard segmentation models. We also observe that the segmentation results obtained by our framework gave a significant improvement for small classes such as poles, person, and rider, which are essential and valuable for autonomous navigation based applications.

****

A "Network Pruning Network'' Approach to Deep Model Compression

Vinay Kumar Verma, Pravendra Singh, Vinay Namboodri, Piyush Rai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3009-3018

We present a filter pruning approach for deep model compression, using a multitask network. Our approach is based on learning a a pruner network to prune a pre-trained target network. The pruner is essentially a multitask deep neural network with binary outputs that help identify the filters from each layer of the original network that do not have any significant contribution to the model and can therefore be pruned. The pruner network has the same architecture as the original network except that it has a multitask/multi-output last layer containing binary-valued outputs (one per filter), which indicate which filters have to be pruned. The pruner's goal is to minimize the number of filters from the original network by assigning zero weights to the corresponding output feature-maps. In contrast to most of the existing methods, instead of relying on iterative pruning, our approach can prune the network (original network) in one go and, moreover, does not require specifying the degree of pruning for each layer (and can learn it instead). The compressed model produced by our approach is generic and does not need any special hardware/software support. Moreover, augmenting with other methods such as knowledge distillation, quantization, and connection pruning can increase the degree of compression for the proposed approach. We show the efficacy of our proposed approach for classification and object detection tasks.

****

GradMix: Multi-source Transfer across Domains and Tasks

Junnan Li, Ziwei Xu, Wong Yongkang, Qi Zhao, Mohan Kankanhalli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3019-3027

The computer vision community is witnessing an unprecedented rate of new tasks being proposed and addressed, thanks to the deep convolutional networks' capability to find complex mappings from X to Y. The advent of each task often accompanies the release of a large-scale annotated dataset, for supervised training of deep network. However, it is expensive and time-consuming to manually label sufficient amount of training data. Therefore, it is important to develop algorithms that can leverage off-the-shelf labeled dataset to learn useful knowledge for the target task. While previous works mostly focus on transfer learning from a single source, we study multi-source transfer across domains and tasks (MS-DTT), in a semi-supervised setting. We propose GradMix, a model-agnostic method applicable to any model trained with gradient-based learning rule, to transfer knowledge via gradient descent by weighting and mixing the gradients from all sources during training. GradMix follows a meta-learning objective, which assigns layer-wise weights to the source gradients, such that the combined gradient follows the direction that minimize the loss for a small set of samples from the target dataset. In addition, we propose to adaptively adjust the learning rate for each mini-batch based on its importance to the target task, and a pseudo-labeling method to leverage the unlabeled samples in the target domain. We conduct MS-DTT experiments on two tasks: digit recognition and action recognition, and demonstrate the advantageous performance of the proposed method against multiple baselines.

****

Jointly Trained Image and Video Generation using Residual Vectors

Yatin Dandi, Aniket Das, Soumye Singhal, Vinay Namboodiri, Piyush Rai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3028-3042

In this work, we propose a modeling technique for jointly training image and video generation models by simultaneously learning to map latent variables with a fixed prior onto real images and interpolate over images to generate videos. The proposed approach models the variations in representations using residual vectors encoding the change at each time step over a summary vector for the entire video. We utilize the technique to jointly train an image generation model with a f

ixed prior along with a video generation model lacking constraints such as disentanglement. The joint training enables the image generator to exploit temporal information while the video generation model learns to flexibly share information across frames. Moreover, experimental results verify our approach's compatibility with pre-training on videos or images and training on datasets containing a mixture of both. A comprehensive set of quantitative and qualitative evaluations reveal the improvements in sample quality and diversity over both video generation and image generation baselines. We further demonstrate the technique's capabilities of exploiting similarity in features across frames by applying it to a model based on decomposing the video into motion and content. The proposed model allows minor variations in content across frames while maintaining the temporal dependence through latent vectors encoding the pose or motion features.
****
CrossNet: Latent Cross-Consistency for Unpaired Image Translation
Omry Sendik, Danny Cohen-Or, Dani Lischinski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3043-3051
Recent GAN-based architectures have been able to deliver impressive performance on the general task of image-to-image translation. In particular, it was shown that a wide variety of image translation operators may be learned from two image sets, containing images from two different domains, without establishing an explicit pairing between the images. This was made possible by introducing clever regularizers to overcome the under-constrained nature of the unpaired translation problem. In this work, we introduce a novel architecture for unpaired image translation, and explore several new regularizers enabled by it. Specifically, our architecture comprises a pair of GANs, as well as a pair of translators between their respective latent spaces. These cross-translators enable us to impose several regularizing constraints on the learnt image translation operator, collectively referred to as latent cross-consistency. Our results show that our proposed architecture and latent cross-consistency constraints are able to outperform the existing state-of-the-art on a variety of image translation tasks.
****
Partially Zero-shot Domain Adaptation from Incomplete Target Data with Missing Classes
Masato Ishii, Takashi Takenouchi, Masashi Sugiyama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3052-3060
We tackle a domain adaptation problem under partially zero-shot setting. In this setting, a certain subset of classes is missing in the unlabeled target data, while all classes appear in the labeled source data, and the goal is to discriminate all classes at the target domain. To solve this problem, we utilize an adversarial training scheme and adopt instance weighting to estimate the loss related to unavailable target data in the missing classes. The instance weight is computed on the basis of the prediction of deep neural networks, implying which instance would be similar to unseen data and having useful information for the loss estimation. This estimation makes it possible to explicitly consider all classes during the domain adaptation training even in the partially zero-shot setting, which leads to accurate adaptation between domains. Experimental results with several benchmark datasets validate the advantage of our method
****
microbatchGAN: Stimulating Diversity with Multi-Adversarial Discrimination
Goncalo Mordido, Haojin Yang, Christoph Meinel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3061-3070
We propose to tackle the mode collapse problem in generative adversarial networks (GANs) by using multiple discriminators and assigning a different portion of each minibatch, called microbatch, to each discriminator. We gradually change each discriminator's task from distinguishing between real and fake samples to discriminating samples coming from inside or outside its assigned microbatch by using a diversity parameter \alpha. The generator is then forced to promote variety in each minibatch to make the microbatch discrimination harder to achieve by each discriminator. Thus, all models in our framework benefit from having variety i

n the generated set to reduce their respective losses. We show evidence that our solution promotes sample diversity since early training stages on multiple data sets.
****

Adversarial Sampling for Active Learning
Christoph Mayer, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3071-3079

This paper proposes ASAL, a new GAN based active learning method that generates high entropy samples. Instead of directly annotating the synthetic samples, ASAL searches similar samples from the pool and includes them for training. Hence, the quality of new samples is high and annotations are reliable. To the best of our knowledge, ASAL is the first GAN based AL method applicable to multi-class problems that outperforms random sample selection. Another benefit of ASAL is its small run-time complexity(sub-linear) compared to traditional uncertainty sampling (linear). We present a comprehensive set of experiments on multiple traditional data sets and show that ASAL outperforms similar methods and clearly exceeds the established baseline (random sampling). In the discussion section we analyze in which situations ASAL performs best and why it is sometimes hard to outperform random sample selection.
****

Resisting Large Data Variations via Introspective Transformation Network
Yunhan Zhao, Ye Tian, Charless Fowlkes, Wei Shen, Alan Yuille; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3080-3089

Training deep networks that generalize to a wide range of variations in test data is essential to building accurate and robust image classifiers. Data variations in this paper include but not limited to unseen affine transformations and warping in the training data. One standard strategy to overcome this problem is to apply data augmentation to synthetically enlarge the training set. However, data augmentation is essentially a brute-force method which generates uniform samples from some pre-defined set of transformations. In this paper, we propose a principled approach named introspective transformation network (ITN) that significantly improves network resistance to large variations between training and testing data. This is achieved by embedding a learnable transformation module into the introspective network, which is a convolutional neural network (CNN) classifier empowered with generative capabilities. Our approach alternates between synthesizing pseudo-negative samples and transformed positive examples based on the current model, and optimizing model predictions on these synthesized samples. Experimental results verify that our approach significantly improves the ability of deep networks to resist large variations between training and testing data and achieves classification accuracy improvements on several benchmark datasets, including MNIST, affNIST, SVHN, CIFAR-10 and miniImageNet.
****

Uncertainty in Model-Agnostic Meta-Learning using Variational Inference
Cuong Nguyen, Thanh-Toan Do, Gustavo Carneiro; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3090-3100

We introduce a new, rigorously-formulated Bayesian meta-learning algorithm that learns a probability distribution of model parameter prior for few-shot learning. The proposed algorithm employs a gradient-based variational inference to infer the posterior of model parameters for a new task. Our algorithm can be applied to any model architecture and can be implemented in various machine learning paradigms, including regression and classification. We show that the models trained with our proposed meta-learning algorithm are well calibrated and accurate, with state-of-the-art calibration and classification results on three few-shot classification benchmarks (Omniglot, mini-ImageNet and tiered-ImageNet), and competitive results in a multi-modal task-distribution regression.
****

A Generative Framework for Zero Shot Learning with Adversarial Domain Adaptation
Varun Khare, Divyat Mahajan, Homanga Bharadhwaj, Vinay Kumar Verma, Piyush Rai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vi

sion (WACV), 2020, pp. 3101-3110

We present a domain adaptation based generative framework for zero shot learning. We address the problem of domain shift between the seen and unseen class distribution in Zero-Shot Learning (ZSL) and seek to minimize it by developing a generative model and training it via adversarial domain adaptation. Our approach is based on end-to-end learning of the class distributions of seen classes and unseen classes. To enable the model to learn the class distributions of unseen classes, we parameterize these class distributions in terms of the class attribute information (which is available for both seen and unseen classes). This provides a very simple way to learn the class distribution of any unseen class, given only its class attribute information, and no labeled training data. Training this model with adversarial domain adaptation provides robustness against the distribution mismatch between the data from seen and unseen classes. It also engenders a novel way for training neural net based classifiers to overcome the hubness problem in Zero-Shot learning. Through a comprehensive set of experiments, we show that our model yields superior accuracies as compared to various state-of-the-art zero shot learning models, on a variety of benchmark datasets.

****

Deep Adaptive Wavelet Network

Maria Ximena Bastidas Rodriguez, Adrien Gruson, Luisa Polania, Shin Fujieda, Flavio Prieto, Kohei Takayama, Toshiya Hachisuka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3111-3119

Even though convolutional neural networks have become the method of choice in many fields of computer vision, they still lack interpretability and are usually designed manually in a cumbersome trial-and-error process. This paper aims at overcoming those limitations by proposing a deep neural network, which is designed in a systematic fashion and is interpretable, by integrating multiresolution analysis at the core of the deep neural network design. By using the lifting scheme, it is possible to generate a wavelet representation and design a network capable of learning wavelet coefficients in an end-to-end form. Compared to state-of-the-art architectures, the proposed model requires less hyper-parameter tuning and achieves competitive accuracy in image classification tasks. The Code implemented for this research is available at https://github.com/mxbastidasr/DAWN_WACV2020

****

Towards Photographic Image Manipulation with Balanced Growing of Generative Autoencoders

Ari Heljakka, Arno Solin, Juho Kannala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3120-3129

We present a generative autoencoder that provides fast encoding, faithful reconstructions (e.g. retaining the identity of a face), sharp generated/reconstructed samples in high resolutions, and a well-structured latent space that supports semantic manipulation of the inputs. There are no current autoencoder or GAN models that satisfactorily achieve all of these. We build on the progressively growing autoencoder model PIONEER, for which we completely alter the training dynamics based on a careful analysis of recently introduced normalization schemes. We show significantly improved visual and quantitative results for face identity conservation in CelebA-HQ. Our model achieves state-of-the-art disentanglement of latent space, both quantitatively and via realistic image attribute manipulations. On the LSUN Bedrooms dataset, we improve the disentanglement performance of the vanilla PIONEER, despite having a simpler model. Overall, our results indicate that the PIONEER networks provide a way towards photorealistic face manipulation.

****

Generative Pseudo-label Refinement for Unsupervised Domain Adaptation

Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, Vittorio Murino; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3130-3139

We investigate and characterize the inherent resilience of conditional Generativ e Adversarial Networks (cGANs) against noise in their conditioning labels, and exploit this fact in the context of Unsupervised Domain Adaptation (UDA). In UDA, a classifier trained on the labelled source set can be used to infer pseudo-labels on the unlabelled target set. However, this will result in a significant amount of misclassified examples (due to the well-known domain shift issue), which can be interpreted as noise injection in the ground-truth labels for the target set. We show that cGANs are, to some extent, robust against such "shift noise". Indeed, cGANs trained with noisy pseudo-labels, are able to filter such noise and generate cleaner target samples. We exploit this finding in an iterative procedure where a generative model and a classifier are jointly trained: in turn, the generator allows to sample cleaner data from the target distribution, and the classifier allows to associate better labels to target samples, progressively refining target pseudo-labels. Results on common benchmarks show that our method performs better or comparably with the unsupervised domain adaptation state of the art.

****

Filter Distillation for Network Compression

Xavier Suau Cuadros, Luca Zappella, Nicholas Apostoloff; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3140-3149

In this paper we introduce Principal Filter Analysis (PFA), an easy to use and effective method for neural network compression. PFA exploits the correlation between filter responses within network layers to recommend a smaller network that maintain as much as possible the accuracy of the full model. We propose two algorithms: the first allows users to target compression to specific network property, such as number of trainable variable (footprint), and produces a compressed model that satisfies the requested property while preserving the maximum amount of spectral energy in the responses of each layer, while the second is a parameter-free heuristic that selects the compression used at each layer by trying to mimic an ideal set of uncorrelated responses. Since PFA compresses networks based on the correlation of their responses we show in our experiments that it gains the additional flexibility of adapting each architecture to a specific domain while compressing. PFA is evaluated against several architectures and datasets, and shows considerable compression rates without compromising accuracy, e.g., for VGG-16 on CIFAR-10, CIFAR-100 and ImageNet, PFA achieves a compression rate of 8x, 3x, and 1.4x with an accuracy gain of 0.4%, 1.4% points, and 2.4% respectively. Our tests show that PFA is competitive with state-of-the-art approaches while removing adoption barriers thanks to its practical implementation, intuitive philosophy and ease of use.

****

How Much Deep Learning does Neural Style Transfer Really Need? An Ablation Study

Len Du; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3150-3159

Neural style transfer has been a "killer app" for deep learning, drawing attention from and advertising the effectiveness to both the academic and the general public. However, we have found by ablative experiments that optimizing an image in the way neural style transfer does, while the objective functions (or more precisely, the functions to transform raw images to corresponding feature maps being compared) are constructed without pretrained weights or biases, worked al- most as well. We can even factor out the deepness (multiple layers of alternating linear and nonlinear transformations) alltogether and have neural style transfer work to a certain extent. This raises the question how much of the the current success of deep learning in computer vision should be attributed to training, structure or simply spatially aggregating the image.

****

Improving Style Transfer with Calibrated Metrics

Mao-Chuang Yeh, Shuai Tang, Anand Bhattad, Chuhang Zou, David Forsyth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3160-3168

Style transfer methods produce a transferred image which is a rendering of a con

tent image in the manner of a style image. We seek to understand how to improve style transfer. To do so requires quantitative evaluation procedures, but current evaluation is qualitative, mostly involving user studies. We describe a novel quantitative evaluation procedure. Our procedure relies on two statistics: the Effectiveness (E) statistic measures the extent that a given style has been transferred to the target, and the Coherence (C) statistic measures the extent to which the original image's content is preserved. Our statistics are calibrated to human preference: targets with larger values of E (resp C) will reliably be preferred by human subjects in comparisons of style (resp. content). We use these statistics to investigate relative performance of a number of style transfer methods, revealing a number of intriguing properties. Admissible methods lie on a Pareto frontier (i.e. improving E reduces C, or vice versa). Three methods are admissible: Universal style transfer produces very good C but weak E; modifying the optimization used for Gatys' loss produces a method with strong E and strong C; and a modified cross-layer method has slightly better E at strong cost in C. While the histogram loss improves the E statistics of Gatys' method, it does not make the method admissible. Surprisingly, style weights have relatively little effect in improving EC scores, and most variability in transfer is explained by the style itself (meaning experimenters can be misguided by selecting styles).
****

Learning from Noisy Labels via Discrepant Collaborative Training
Yan Han, SOUMAVA ROY, Lars Petersson, Mehrtash Harandi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3169-3178
Noise is ubiquitous in the world around us. Difficulty inestimating the noise within a dataset makes learning fromsuch a dataset a difficult and challenging task. In this pa-per, we propose a novel and effective learning frameworkin order to alleviate the adverse effects of noise within adataset. Towards this aim, we modify a collaborative train-ing framework to utilize discrepancy constraints betweenrespective feature extractors enabling the learning of distinct, yet discriminative features, pacifying the adverse ef-fects of noise. Empirical results of our proposed algo-rithm, Discrepant Collaborative Training (DCT), achievecompetitive results against several current state-of-the-artalgorithms across MNIST, CIFAR10 and CIFAR100, as wellas large fine-grained image classification datasets such asCUBS-200-2011 and CARS196 for different levels of noise.
****

Class-Discriminative Feature Embedding For Meta-Learning based Few-Shot Classification
Alireza Rahimpour, Hairong Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3179-3187
Although deep learning-based approaches have been very effective in solving problems with plenty of labeled data, they suffer in tackling problems for which labeled data are scarce. In few-shot classification, the objective is to train a classifier from only a handful of labeled examples in a support set. In this paper, we propose a few-shot learning framework based on structured margin loss which takes into account the global structure of the support set in order to generate a highly discriminative feature space where the features from distinct classes are well separated in clusters. Moreover, in our meta-learning-based framework, we propose a context-aware query embedding encoder for incorporating support set context into query embedding and generating more discriminative and task-dependent query embeddings. The task-dependent features help the meta-learner to learn a distribution over tasks more effectively. Extensive experiments based on few-shot, zero-shot and semi-supervised learning on three benchmarks show the advantages of the proposed model compared to the state-of-the-art.
****

Adaptive Neural Connections for Sparsity Learning
Alex Gain, Prakhar Kaushik, Hava Siegelmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3188-3193
Sparsity learning aims to decrease the computational and memory costs of large d

eep neural networks (DNNs) via pruning neural connections while simulaneously retaining high accuracy. A large body of work has developed sparsity learning approaches, with recent large-scale experiments showing that two main methods, magnitude pruning and Variational Dropout (VD), achieve similar state-of-the-art results for classification tasks. We propose Adaptive Neural Connections (ANC), a method for explicitly parameterizing fine-grained neuron-to-neuron connections via adjacency matrices at each layer that are learned through backpropagation. Explicitly parameterizing neuron-to-neuron connections confers two primary advantages: 1. Sparsity can be explicitly optimized for via norm-based regularization on the adjacency matrices; and 2. When combined with VD (which we term, ANC-VD), the adjacencies can be interpreted as learned weight importance parameters, which we hypothesize leads to improved convergence for VD. Experiments with ResNet18 show that architectures augmented with ANC outperform their vanilla counterparts.

****

FX-GAN: Self-Supervised GAN Learning via Feature Exchange

Rui Huang, Wenju Xu, Teng-Yok Lee, Anoop Cherian, Ye Wang, Tim Marks; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3194-3202

We propose a self-supervised approach to improve the training of Generative Adversarial Networks (GANs) via inducing the discriminator to examine the structural consistency of images. Although natural image samples provide ideal examples of both valid structure and valid texture, learning to reproduce both together remains an open challenge. In our approach, we augment the training set of natural images with modified examples that have degraded structural consistency. These degraded examples are automatically created by randomly exchanging pairs of patches in an image's convolutional feature map. We call this approach feature exchange. With this setup, we propose a novel GAN formulation, termed Feature eXchange GAN (FX-GAN), in which the discriminator is trained not only to distinguish real versus generated images, but also to perform the auxiliary task of distinguishing between real images and structurally corrupted (feature-exchanged) real images. This auxiliary task causes the discriminator to learn the proper feature structure of natural images, which in turn guides the generator to produce images with more realistic structure. Compared with strong GAN baselines, our proposed self-supervision approach improves generated image quality, diversity, and training stability for both the unconditional and class-conditional settings.

****

Synthesizing human-like sketches from natural images using a conditional convolutional decoder

Moritz Kampelmuhler, Axel Pinz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3203-3211

Humans are able to precisely communicate diverse concepts by employing sketches, a highly reduced and abstract shape based representation of visual content. We propose, for the first time, a fully convolutional end-to-end architecture that is able to synthesize human-like sketches of objects in natural images with potentially cluttered background. To enable an architecture to learn this highly abstract mapping, we employ the following key components: (1) a fully convolutional encoder-decoder structure, (2) a perceptual similarity loss function operating in an abstract feature space and (3) conditioning of the decoder on the label of the object that shall be sketched. Given the combination of these architectural concepts, we can train our structure in an end-to-end supervised fashion on a collection of sketch-image pairs. The generated sketches of our architecture can be classified with 85.6% Top-5 accuracy and we verify their visual quality via a user study. We find that deep features as a perceptual similarity metric enable image translation with large domain gaps and our findings further show that convolutional neural networks trained on image classification tasks implicitly learn to encode shape information.

****

Best Frame Selection in a Short Video

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3212-3221

People usually take short videos to record meaningful moments in their lives. However, selecting the most representative frame, which not only has high image visual quality but also captures video content, from a short video to share or keep is a time-consuming process for one may need to manually go through all the frames in a video to make a decision. In this paper, we introduce the problem of the best frame selection in a short video and aim to solve it automatically. Towards this end, we collect and will release a diverse large-scale short video dataset that includes 11, 000 videos shoot in our daily life. All videos are assumed to be short (e.g., a few seconds) and each video has human-annotated of the best frame. Then we introduce a deep convolutional neural network (CNN) based approach with ranking objective to automatically pick the best frame from frame sequences extracted via short videos. Additionally, we propose new evaluation metrics, especially for the best frame selection. In experiments, we show our approach outperforms various other methods significantly.
****

Fast Video Multi-Style Transfer
Wei Gao, Yijun Li, Yihang Yin, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3222-3230

Recent progresses in video style transfer have shown promising results which contain less flickering effects. However, existing algorithms mainly trade off generality for efficiency, i.e., constructing one network per style example, and often work well for short video clips only. Specifically, we design a multi-instance normalization block (MIN-Block) to learn different style examples and a ConvLSTM module to encourage the temporal consistency. The proposed algorithm is demonstrated to be able to generate temporally-consistent video transfer results in different styles while keeping each stylized frame visually pleasing. Extensive experimental results show that the proposed method performs favorably again single-style models and some post-processing techniques that alleviate the flickering issue. We achieve as many as 120 stylization effects in a single model and show results on long-term videos that consist of thousands of frames.
****

Toward Interactive Self-Annotation For Video Object Bounding Box: Recurrent Self-Learning And Hierarchical Annotation Based Framework
Trung-Nghia Le, Akihiro Sugimoto, Shintaro Ono, Hiroshi Kawasaki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3231-3240

Amount and variety of training data drastically affect the performance of CNNs. Thus, annotation methods are becoming more and more critical to collect data efficiently. In this paper, we propose a simple yet efficient Interactive Self-Annotation framework to cut down both time and human labor cost for video object bounding box annotation. Our method is based on recurrent self-supervised learning and consists of two processes: automatic process and interactive process, where the automatic process aims to build a supported detector to speed up the interactive process. In the Automatic Recurrent Annotation, we let an off-the-shelf detector watch unlabeled videos repeatedly to reinforce itself automatically. At each iteration, we utilize the trained model from the previous iteration to generate better pseudo ground-truth bounding boxes than those at the previous iteration, recurrently improving self-supervised training the detector. In the Interactive Recurrent Annotation, we tackle the human-in-the-loop annotation scenario where the detector receives feedback from the human annotator. To this end, we propose a novel Hierarchical Correction module, where the annotated frame-distance binarizedly decreases at each time step, to utilize the strength of CNN for neighbor frames. Experimental results on various video datasets demonstrate the advantages of the proposed framework in generating high-quality annotations while reducing annotation time and human labor costs.
****

TailorGAN: Making User-Defined Fashion Designs
Lele Chen, Justin Tian, Guo Li, Cheng-Haw Wu, Erh-Kan King, Kuan-Ting Chen, Shao-Hang Hsieh, Chenliang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3241-3250

Attribute editing has become an important and emerging topic of computer vision. In this paper, we consider a task: given a reference garment image A and another image B with target attribute (collar/sleeve), generate a photo-realistic image which combines the texture from reference A and the new attribute from reference B. The highly convoluted attributes and the lack of paired data are the main challenges to the task. To overcome those limitations, we propose a novel self-supervised model to synthesize garment images with disentangled attributes (e.g., collar and sleeves) without paired data. Our method consists of reconstruction learning step and adversarial learning step. The model learns texture and location information through reconstruction learning. And the model capability is generalized to achieve single-attribute manipulation by adversarial learning. Meanwhile, we compose a new dataset, named GarmentSet, with annotation of landmarks of collar and sleeves on clean garment images. Thoughtful experiments on this dataset and real-world samples demonstrate that our method can synthesize significantly better results than the state-of-the-art methods in both quantitative and qualitative comparisons. The code is available at: https://github.com/gli-27/TailorGAN.

****

Coordinated Joint Multimodal Embeddings for Generalized Audio-Visual Zero-shot Classification and Retrieval of Videos

Kranti Parida, Neeraj Matiyali, Tanaya Guha, Gaurav Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3251-3260

We present an audio-visual multimodal approach for the task of zero-shot learning (ZSL) for classification and retrieval of videos. ZSL has been studied extensively in the recent past but has primarily been limited to visual modality and to images. We demonstrate that both audio and visual modalities are important for ZSL for videos. Since a dataset to study the task is currently not available, we also construct an appropriate multimodal dataset with 33 classes containing 156, 416 videos, from an existing large scale audio event dataset. We empirically show that the performance improves by adding audio modality for both tasks of zero-shot classification and retrieval, when using multi-modal extensions of embedding learning methods. We also propose a novel method to predict the 'dominant' modality using a jointly learned modality attention network. We learn the attention in a semi-supervised setting and thus do not require any additional explicit labelling for the modalities. We provide qualitative validation of the modality specific attention, which also successfully generalizes to unseen test classes.

****

s-SBIR: Style Augmented Sketch based Image Retrieval

Titir Dutta, Soma Biswas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3261-3270

Sketch-based image retrieval (SBIR) is gaining increasing popularity because of its flexibility to search natural images using unrestricted hand-drawn sketch query. Here, we address a related, but relatively unexplored problem, where the users can also specify their preferred styles of the images they want to retrieve, e.g., color, shape, etc., as key-words, whose information is not present in the sketch. The contribution of this work is three-fold. First, we propose a deep network for the problem of style-augmented SBIR (or s-SBIR) having three main components - category module, style module and mixer module, which are trained in an end-to-end manner. Second, we propose a quintuplet loss, which takes into consideration both the category and style, while giving appropriate importance to the two components. Third, we propose a composite evaluation metric or ncMAP which can quantitatively evaluate s-SBIR approaches. Extensive experiments on subsets of two benchmark image-sketch datasets, Sketchy and TU-Berlin show the effectiveness of the proposed approach.

****

Personalizing Fast-Forward Videos Based on Visual and Textual Features from Social Network

Washington Ramos, Michel Silva, Edson Araujo, Alan Neves, Erickson Nascimento; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vis

ion (WACV), 2020, pp. 3271-3280
The growth of Social Networks has fueled the habit of people logging their day-to-day activities, and long First-Person Videos (FPVs) are one of the main tools in this new habit. Semantic-aware fast-forward methods are able to decrease the watch time and select meaningful moments, which is key to increase the chances of these videos being watched. However, these methods can not handle semantics in terms of personalization. In this paper, we present a new approach to automatically creating personalized fast-forward videos for FPVs. Our approach explores the availability of text-centric data from the user's social networks such as status updates to infer her/his topics of interest and assigns scores to the input frames according to her/his preferences. Extensive experiments are conducted on three different datasets with simulated and real-world users as input. Our method achieved an average F1 score of up to 12.8 percentage points higher than the best competitors. We also present a user study to demonstrate the effectiveness of our method.
****

Unsupervised Image Style Embeddings for Retrieval and Recognition Tasks
Siddhartha Gairola, Rajvi Shah, P. J. Narayanan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3281-3289
We propose an unsupervised protocol for learning a neural embedding of visual style of images. Style similarity is an important measure for many applications such as style transfer, fashion search, art exploration, etc. However, computational modeling of style is a difficult task owing to its vague and subjective nature. Most methods for style based retrieval use supervised training with pre-defined categorization of images according to style. While this paradigm is suitable for applications where style categories are well-defined and curating large datasets according to such a categorization is feasible, in several other cases such a categorization is either ill-defined or does not exist. Our protocol for learning style based representations does not leverage categorical labels but a proxy measure for forming triplets of anchor, similar, and dissimilar images. Using these triplets, we learn a compact style embedding that is useful for style-based search and retrieval. The learned embeddings outperform other unsupervised representations for style-based image retrieval task on six datasets that capture different meanings of style. We also show that by fine-tuning the learned features with dataset-specific style labels, we obtain best results for image style recognition task on five of six datasets.
****

Animating Face using Disentangled Audio Representations
Gaurav Mittal, Baoyuan Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3290-3298
Previous methods for audio-driven talking head generation assume the input audio to be clean with a neutral tone. As we show empirically, one can easily break these systems by simply adding certain background noise to the utterance or changing its emotional tone (to for example, sad). To make talking head generation robust to such variations, we propose an explicit audio representation learning framework that disentangles audio sequences into various factors such as phonetic content, emotional tone, background noise and others. We conduct experiments to validate that when conditioned on disentangled content representation, the generated mouth movement by our model is significantly more accurate than previous approaches (without disentangled learning) in the presence of noise and emotional variations. We further demonstrate that our framework is compatible with current state-of-the-art approaches by replacing their original component to learn audio based representation with ours. To the best of our knowledge, this is the first work which improves the performance of talking head generation through a disentangled audio representation perspective, which is important for many real-world applications.
****

High-Frequency Refinement for Sharper Video Super-Resolution
Vikram Singh, Akshay Sharma, Sudharshann Devanathan, Anurag Mittal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV),

2020, pp. 3299-3308

A video super-resolution technique is expected to generate a `sharp' upsampled video. The sharpness in the generated video comes from the precise prediction of the high-frequency details (e.g. object edges). Thus high-frequency prediction becomes a vital sub-problem of the super-resolution task. To generate a sharp-upsampled video, this paper proposes an upsampling network architecture `HFR-Net' that works on the principle of `explicit refinement and fusion of high-frequency details'. To implement this principle and to train HFR-Net, a novel technique named 2-phase progressive-retrogressive training is being proposed. Additionally, a method called dual motion warping is also being introduced to preprocess the videos that have varying motion intensities (slow and fast). Results on multiple video datasets demonstrate the improved performance of our approach over the current state-of-the-art.

****

AlignNet: A Unifying Approach to Audio-Visual Alignment

Jianren Wang, Zhaoyuan Fang, Hang Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3309-3317

We present AlignNet, a model that synchronizes videos with reference audios under non-uniform and irregular misalignments. AlignNet learns the end-to-end dense correspondence between each frame of a video and an audio. Our method is designed according to simple and well-established principles: attention, pyramidal processing, warping, and affinity function. Together with the model, we release a dancing dataset Dance50 for training and evaluation. Qualitative, quantitative and subjective evaluation results on dance-music alignment and speech-lip alignment demonstrate that our method far outperforms the state-of-the-art methods. Code, dataset and sample videos are available at our project page.

****

Eye Contact Correction using Deep Neural Networks

Furkan Isikdogan, Timo Gerasimow, Gilad Michael; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3318-3326

In a typical video conferencing setup, it is hard to maintain eye contact during a call since it requires looking into the camera rather than the display. We propose an eye contact correction model that restores the eye contact regardless of the relative position of the camera and display. Unlike previous solutions, our model redirects the gaze from an arbitrary direction to the center without requiring a redirection angle or camera/display/user geometry as inputs. We use a deep convolutional neural network that inputs a monocular image and produces a vector field and a brightness map to correct the gaze. We train this model in a bi-directional way on a large set of synthetically generated photorealistic images with perfect labels. The learned model is a robust eye contact corrector which also predicts the input gaze implicitly at no additional cost.   Our system is primarily designed to improve the quality of video conferencing experience. Therefore, we use a set of control mechanisms to prevent creepy results and to ensure a smooth and natural video conferencing experience. The entire eye contact correction system runs end-to-end in real-time on a commodity CPU and does not require any dedicated hardware, making our solution feasible for a variety of devices.

****

Attention Flow: End-to-End Joint Attention Estimation

Omer Sumer, Peter Gerjets, Ulrich Trautwein, Enkelejda Kasneci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3327-3336

This paper addresses the problem of understanding joint attention in third-person social scene videos. Joint attention is the shared gaze behaviour of two or more individuals on an object or an area of interest and has a wide range of applications such as human-computer interaction, educational assessment, treatment of patients with attention disorders, and many more. Our method, Attention Flow, learns joint attention in an end-to-end fashion by using saliency-augmented attention maps and two novel convolutional attention mechanisms that determine to select relevant features and improve joint attention localization. We compare the e ffect of saliency maps and attention mechanisms and report quantitative and qualitative results on the detection and localization of joint attention in the VideoCoAtt dataset, which contains complex social scenes.

****

PSNet: A Style Transfer Network for Point Cloud Stylization on Geometry and Color

Xu Cao, Weimin Wang, Katashi Nagao, Ryosuke Nakamura; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3337-3345

We propose a neural style transfer method for colored point clouds which allows stylizing the geometry and/or color property of a point cloud from another. The stylization is achieved by manipulating the content representations and Gram-based style representations extracted from a pre-trained PointNet-based classification network for colored point clouds. As Gram-based style representation is invariant to the number or the order of points, the style can also be an image in the case of stylizing the color property of a point cloud by merely treating the image as a set of pixels. Experimental results and analysis demonstrate the capability of the proposed method for stylizing a point cloud either from another point cloud or an image.

****

Neural Puppet: Generative Layered Cartoon Characters

Omid Poursaeed, Vladimir Kim, Eli Shechtman, Jun Saito, Serge Belongie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3346-3356

We propose a learning based method for generating new animations of a cartoon character given a few example images. Our method is designed to learn from a traditional animation, where each frame is drawn by an artist, and thus the input images lack any common structure, correspondences, or labels. We express pose changes as a deformation of a layered 2.5D template mesh, and devise a novel architecture that learns to predict mesh deformations matching the template to a target image. This enables us to extract a common low-dimensional structure in the diverse set of character poses. We combine recent advances in differentiable rendering as well as mesh-aware models to successfully align common template even if only a few character images are available during training. In addition to coarse poses, character appearance also varies due to shading, out-of-plane motions, and artistic effects. We capture these subtle changes by applying an image translation network to refine the mesh rendering, providing an end-to-end model to generate new animations of a character with high visual quality. We demonstrate that our generative model can be used to synthesize in-between frames and to create data-driven deformation. Our template fitting procedure outperforms state-of-the-art generic techniques for detecting image correspondences.

****

BRDF-Reconstruction in Photogrammetry Studio Setups

Matthias Innmann, Jochen Sussmuth, Marc Stamminger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3357-3365

Photogrammetry Studios are a common setup to acquire high-quality 3D geometry from different kinds of real-world objects, humans, etc. In a photo studio like setup, 50 - 200 DSLR cameras are used with object-specific illumination to simultaneously capture images that are processed by algorithms that automatically estimate the camera parameters and detailed geometry. These steps are automated in established pipelines to a large extent and do not require much user input. However, the post-processing typically involves a manual estimation of surface reflectance parameters by an artist, who paints textures to allow for photorealistic rendering. While professional light stages facilitate this process in an automated way, these setups are very expensive and require accurately calibrated light sources and cameras. In our work, we present a new formulation along with a practical solution to reduce these constraints to photo studio like setups by jointly reconstructing the geometric configuration of the lights along with spatially varying surface reflectance properties and its diffuse albedo. In the presented sy nthetic as well as real-world experiments, we analyze the effect of different optimization objectives and show that our method is able to provide photorealistic reconstruction results with an RMSE of 1 - 3 % on real data.

****

Do As I Do: Transferring Human Motion and Appearance between Monocular Videos with Spatial and Temporal Constraints

Thiago Gomes, Renato Martins, Joao Ferreira, Erickson Nascimento; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3366-3375

Creating plausible virtual actors from images of real actors remains one of the key challenges in computer vision and computer graphics. Marker-less human motion estimation and shape modeling from images in the wild bring this challenge to the fore. Although the recent advances on view synthesis and image-to-image translation, currently available formulations are limited to transfer solely style and do not take into account the character's motion and shape, which are by nature intermingled to produce plausible human forms. In this paper, we propose a unifying formulation for transferring appearance and retargeting human motion from monocular videos that regards all these aspects. Our method synthesizes new videos of people in a different context where they were initially recorded. Differently from recent appearance transferring methods, our approach takes into account body shape, appearance, and motion constraints. The evaluation is performed with several experiments using publicly available real videos containing hard conditions. Our method is able to transfer both human motion and appearance outperforming state-of-the-art methods, while preserving specific features of the motion that must be maintained (e.g., feet touching the floor, hands touching a particular object) and holding the best visual quality and appearance metrics such as Structural Similarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS)

****

Temporal Aggregation with Clip-level Attention for Video-based Person Re-identification

Mengliu Li, Han Xu, Jinjun Wang, Wenpeng Li, Yongli Sun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3376-3384

Video-based person re-identification (Re-ID) methods can extract richer features than image-based ones from short video clips. The existing methods usually apply simple strategies, such as average/max pooling, to obtain the tracklet-level features, which has been proved hard to aggregate the information from all video frames. In this paper, we propose a simple yet effective Temporal Aggregation with Clip-level Attention Network (TACAN) to solve the temporal aggregation problem in a hierarchal way. Specifically, a tracklet is firstly broken into different numbers of clips, through a two-stage temporal aggregation network we can get the tracklet-level feature representation. A novel min-max loss is introduced to learn both a clip-level attention extractor and a clip-level feature representer in the training process. Afterwards, the resulting clip-level weights are further taken to average the clip-level features, which can generate a robust tracklet-level feature representation at the testing stage. Experimental results on four benchmark datasets, including the MARS, iLIDS-VID, PRID-2011 and DukeMTMC-VideoReID, show that our TACAN has achieved significant improvements as compared with the state-of-the-art approaches.

****

ICface: Interpretable and Controllable Face Reenactment Using GANs

Soumya Tripathy, Juho Kannala, Esa Rahtu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3385-3394

This paper presents a generic face animator that can control the pose and expressions of a given face image. The animation is driven by human interpretable control signals consisting of head pose angles and the Action Unit (AU) values. The control information can be obtained from multiple sources including external driving videos and manual controls. Due to the interpretable nature of the driving signal, one can easily mix the information between multiple sources (e.g. pose f rom one image and expression from another) and apply selective post- production editing. The proposed face animator is implemented as a two-stage neural network model that is learned in a self-supervised manner using a large video collection. The proposed Interpretable and Controllable face reenactment network (ICface) is compared to the state-of-the-art neural network-based face animation techniques in multiple tasks. The results indicate that ICface produces better visual quality while being more versatile than most of the comparison methods. The introduced model could provide a lightweight and easy to use tool for a multitude of advanced image and video editing tasks. The program code will be publicly available upon the acceptance of the paper.

****

Neural Sign Language Synthesis: Words Are Our Glosses

Jan Zelinka, Jakub Kanis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3395-3403

This paper deals with a text-to-video sign language synthesis. Instead of direct video production, we focused on skeletal models production. Our main goal in this paper was to design the first fully end-to-end automatic sign language synthesis system trained only on available free data (daily TV broadcasting). Thus, we excluded any manual video annotation. Furthermore, our designed approach even do not rely on any video segmentation. A proposed feed-forward transformer and recurrent transformer were investigated. To improve the performance of our sequence-to-sequence transformer, soft non-monotonic attention was employed in our training process. A benefit of character-level features was compared with word-level features. Besides a novel approach to sign language synthesis, we also present a gradient-descend-based method for the skeletal model estimation improvement. This improvement not only smooths skeletal models and interpolates missing bones but it also creates 3D skeletal models from 2D models. We focused our experiments on a weather forecasting dataset in the Czech Sign Language.

****

Preference-Based Image Generation

Hadi Kazemi, Fariborz Taherkhani, Nasser Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3404-3413

Deep generative models are a set of promising methods, that are able to model complex data and generate new samples. In principle, they learn to map a random latent code sampled from a prior distribution into high dimensional data space, such as image space. However, these models have limited utilities as the user has minimal control over what the network produces. Despite the success of some recent work in learning an interpretable latent code, the field still lacks a coherent framework to learn a fully interpretable latent code, without any random part for sample diversity.   Consequently, it is generally hard, if not impossible, for a non-expert user to produce the desired image by tuning the random and interpretable parts of the latent code. In this paper, we introduce the Preference-Based Image Generation (PbIG), a new method to retrieve the corresponding latent code of the user's mental image. We propose to adopt preference-based reinforcement learning, which learns from a user's judgment of the generated images by a pre-trained generative model. Since the proposed method is completely decoupled from the training stage of the underlying generative models, it can easily be adopted by any method, such as GANs and VAEs. We evaluate the effectiveness of PbIG framework using a set of experiments on baseline datasets using a pretraind StackGAN++.

****

Body Pose Sonification for a View-Independent Auditory Aid to Blind Rock Climbers

Joseph Ramsay, Hyung Jin Chang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3414-3421

Rock climbing is a sport in which blind people have traditionally found it extremely difficult to excel due to the high degree of visual problem solving required, and also the requirement to climb with a sighted assistant. We present a system which automates the role of the sighted assistant in order to provide blind p eople with the freedom to climb and train on their own. We address climbing-specific limitations of a state-of-the-art skeleton tracking system, and discuss the ways in which we mitigated these limitations using post-processing techniques tuned specially for a climbing scenario. We also describe the auditory feedback system used to instruct the blind climber, and demonstrate that a user can learn to follow it in a relatively short time by showing a significant improvement in performance over just a few trials with the system.

****

Hand-Priming in Object Localization for Assistive Egocentric Vision

Kyungjun Lee, Abhinav Shrivastava, Hernisa Kacorri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3422-3432

Egocentric vision holds great promises for increasing access to visual information and improving the quality of life for people with visual impairments, with object recognition being one of the daily challenges for this population. While we strive to improve recognition performance, it remains difficult to identify which object is of interest to the user; the object may not even be included in the frame due to challenges in camera aiming without visual feedback. Also, gaze information, commonly used to infer the area of interest in egocentric vision, is often not dependable. However, blind users often tend to include their hand either interacting with the object that they wish to recognize or simply placing it in proximity for better camera aiming. We propose localization models that leverage the presence of the hand as the contextual information for priming the center area of the object of interest. In our approach, hand segmentation is fed to either the entire localization network or its last convolutional layers. Using egocentric datasets from sighted and blind individuals, we show that the hand-priming achieves higher precision than other approaches, such as fine-tuning, multi-class, and multi-task learning, which also encode hand-object interactions in localization.

****

Multimodal Image Outpainting With Regularized Normalized Diversification

Lingzhi Zhang, Jiancong Wang, Jianbo Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3433-3442

In this paper, we study the problem of generating a set of realistic and diverse backgrounds when given only a small foreground region. We refer to this task as image outpainting. The technical challenge of this task is to synthesize not only plausible but also diverse image outputs. Traditional generative adversarial networks suffer from mode collapse. While recent approaches propose to maximize or preserve the pairwise distance between generated samples with respect to their latent distance, they do not explicitly prevent the diverse samples of different conditional inputs from collapsing. Therefore, we propose a new regularization method to encourage diverse sampling in this conditional synthesis. In addition, we propose a novel feature pyramid discriminator to improve the image quality. Our experimental results show that our model can produce more diverse images without sacrificing visual quality compared to state-of-the-arts approaches in both the CelebA face dataset and the Cityscape scene dataset.

****

Learning to Detect Head Movement in Unconstrained Remote Gaze Estimation in the Wild

Zhecan Wang, Jian Zhao, Cheng Lu, Fan Yang, Han Huang, lianji li, Yandong Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3443-3452

Unconstrained remote gaze estimation remains challenging mostly due to its vulnerability to the large variability in head-pose. Prior solutions struggle to maintain reliable accuracy in unconstrained remote gaze tracking. Among them, appearance-based solutions demonstrate tremendous potential in improving gaze accuracy. However, existing works still suffer from head movement and are not robust enough to handle real-world scenarios. Especially most of them study gaze estimation under controlled scenarios where the collected datasets often cover limited ranges of both head-pose and gaze which introduces further bias. In this paper, we

propose novel end-to-end appearance-based gaze estimation methods that could more robustly incorporate different levels of head-pose representations into gaze estimation. Our method could generalize to real-world scenarios with low image quality, different lightings and scenarios where direct head-pose information is not available. To better demonstrate the advantage of our methods, we further propose a new benchmark dataset with the most rich distribution of head-gaze combination reflecting real-world scenarios. Extensive evaluations on several public datasets and our own dataset demonstrate that our method consistently outperforms the state-of-the-arts by a significant margin.
****

MoBiNet: A Mobile Binary Network for Image Classification
Hai Phan, Dang The Huynh, Yihui He, Marios Savvides, Zhiqiang Shen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3453-3462

MobileNet and Binary Neural Networks are two among the most widely used techniques to construct deep learning models for performing a variety of tasks on mobile and embedded platforms.In this paper, we present a simple yet efficient scheme to exploit MobileNet binarization at activation function and model weights. However, training a binary network from scratch with separable depth-wise and point-wise convolutions in case of MobileNet is not trivial and prone to divergence. To tackle this training issue, we propose a novel neural network architecture, namely MoBiNet - Mobile Binary Network in which skip connections are manipulated to prevent information loss and vanishing gradient, thus facilitate the training process. More importantly, while existing binary neural networks often make use of cumbersome backbones such as Alex-Net, ResNet, VGG-16 with float-type pre-trained weights initialization, our MoBiNet focuses on binarizing the already-compressed neural networks like MobileNet without the need of a pre-trained model to start with. Therefore, our proposal results in an effectively small model while keeping the accuracy comparable to existing ones. Experiments on ImageNet dataset show the potential of the MoBiNet as it achieves 54.40% top-1 accuracy and dramatically reduces the computational cost with binary operators.
****

Image Difficulty Curriculum for Generative Adversarial Networks (CuGAN)
Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, Marius Leordeanu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3463-3472

Despite the significant advances in recent years, Generative Adversarial Networks (GANs) are still notoriously hard to train. In this paper, we propose three novel curriculum learning strategies for training GANs. All strategies are first based on ranking the training images by their difficulty scores, which are estimated by a state-of-the-art image difficulty predictor. Our first strategy is to divide images into gradually more difficult batches. Our second strategy introduces a novel curriculum loss function for the discriminator that takes into account the difficulty scores of the real images. Our third strategy is based on sampling from an evolving distribution, which favors the easier images during the initial training stages and gradually converges to a uniform distribution, in which samples are equally likely, regardless of difficulty. We compare our curriculum learning strategies with the classic training procedure on two tasks: image generation and image translation. Our experiments indicate that all strategies provide faster convergence and superior results. For example, our best curriculum learning strategy applied on spectrally normalized GANs (SNGANs) fooled human annotators in thinking that generated CIFAR-like images are real in 25.0% of the presented cases, while the SNGANs trained using the classic procedure fooled the annotators in only 18.4% cases. Similarly, in image translation, the human annotators preferred the images produced by the Cycle-consistent GAN (CycleGAN) trained using curriculum learning in 40.5% cases and those produced by CycleGAN based on classic training in only 19.8% cases, 39.7% cases being labeled as ties.
****

Fast Postprocessing for Difficult Discrete Energy Minimization Problems
Ijaz Akhter, Loong Fah Cheong, RICHARD HARTLEY; Proceedings of the IEEE/CVF Wi

nter Conference on Applications of Computer Vision (WACV), 2020, pp. 3473-3482

Despite the rapid progress in discrete energy minimization, certain problems involving high connectivity and a high number of labels are considered very hard but are still very relevant in computer vision. We propose a post-processing technique to improve the sub-optimal results of the existing methods on such problems. Our core contribution is a mapping between the binary min-cut problem and finding the shortest path in a directed acyclic graph. Using this mapping, we present an algorithm to find an approximate solution for the min-cut problem. We also extend the same idea for multi-label factor-graphs in the form of an iterative move-making algorithm. The proposed algorithm is extremely fast, yet outperforms the existing techniques in terms of accuracy as well as the computational time. We demonstrate competitive or better results on problems where already high-quality work is done.

****

CoachGAN

Mike Brodie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3483-3492

CoachGAN provides an inference time method to improve outputs from GAN generator models. Similar to creating adversarial examples to fool neural network classifiers, CoachGAN exploits gradient information, in this case from a pretrained discriminator model. Unlike the process of generating adversarial examples, which uses gradient descent to alter outputs directly, CoachGAN alters the inputs of generator models. This allows for output enhancements at test time without any additional model training. CoachGAN adapts easily to existing algorithms and does not depend on specific model architectures. In addition to qualitative samples, we quantitatively demonstrate the ability of CoachGAN to improve IS and FID scores across a variety of GAN architectures and tasks.

****

Towards Preserving the Ephemeral: Texture-Based Background Modelling for Capturing Back-of-the-Napkin Notes

Melissa Cote, Alexandra Branzan Albu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3493-3501

A back-of-the-napkin idea is typically created on the spur of the moment and captured via a few hand-sketched notes on whatever material is available, which often happens to be an actual paper napkin. This paper explores the preservation of such back-of-the-napkin ideas. Hand-sketched notes, reflecting those flashes of inspiration, are not limited to text; they can also include drawings and graphics. Napkin backgrounds typically exhibit diverse textural and colour motifs/patterns that may have high visual saliency from a low-level vision standpoint. We thus frame the extraction of hand-sketched notes as a background modelling and removal task. We propose a novel document background model based on texture mixtures constructed from the document itself via texture synthesis, which allows us to identify background pixels and extract hand-sketched data as foreground elements. Experiments on a novel napkin image dataset yield excellent results and showcase the robustness of our method with respect to the napkin contents. A texture-based background modelling approach, such as ours, is generic enough to cope with any type of hand-sketched notes.

****

Unsupervised Writer Adaptation for Synthetic-to-Real Handwritten Word Recognition

Lei Kang, Marcal Rusinol, Alicia Fornes, Pau Riba, Mauricio Villegas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3502-3511

Handwritten Text Recognition (HTR) is still a challenging problem because it must deal with two important difficulties: the variability among writing styles, and the scarcity of labelled data. To alleviate such problems, synthetic data generation and data augmentation are typically used to train HTR systems. However, training with such data produces encouraging but still inaccurate transcriptions in real words. In this paper, we propose an unsupervised writer adaptation approach that is able to automatically adjust a generic handwritten word recognizer,

fully trained with synthetic fonts, towards a new incoming writer. We have experimentally validated our proposal using five different datasets, covering several challenges (i) the document source: modern and historic samples, which may involve paper degradation problems; (ii) different handwriting styles: single and multiple writer collections; and (iii) language, which involves different character combinations. Across these challenging collections, we show that our system is able to maintain its performance, thus, it provides a practical and generic approach to deal with new document collections without requiring any expensive and tedious manual annotation s
****

LEAF-QA: Locate, Encode & Attend for Figure Question Answering
Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, Ajay Joshi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3512-3521
We introduce LEAF-QA, a comprehensive dataset of 250,000 densely annotated figures/charts, constructed from real-world open data sources, along with 2 million question-answer (QA) pairs querying the structure and semantics of these charts. LEAF-QA highlights the problem of multimodal QA, which is notably different from conventional visual QA (VQA), and has recently gained interest in the community. Furthermore, LEAF-QA is significantly more complex than previous attempts at chart QA, viz. FigureQA and DVQA, which present only limited variations in chart data. LEAF-QA being constructed from real-world sources, requires a novel architecture to enable question answering. To this end, LEAF-Net, a deep architecture involving chart element localization, question and answer encoding in terms of chart elements, and an attention network is proposed. Different experiments are conducted to demonstrate the challenges of QA on LEAF-QA. The proposed architecture, LEAF-Net also considerably advances the current state-of-the-art on FigureQA and DVQA.
****

DeepErase: Weakly Supervised Ink Artifact Removal in Document Text Images
Yike Qi, W. Ronny Huang, Qianqian Li, Jonathan Degange; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3522-3530
Paper-intensive industries like insurance, law, and government have long leveraged optical character recognition (OCR) to automatically transcribe hordes of scanned documents into text strings for downstream processing. Even today, there are still many scanned documents and mail that come into businesses in non-digital format. Text to be extracted from real world documents is often nestled inside rich formatting, such as tabular structures or forms with fill-in-the-blank boxes or underlines whose ink often touches or even strikes through the ink of the text itself. Further, the text region could have random ink smudges or spurious strokes. Such ink artifacts can severely interfere with the performance of recognition algorithms or other downstream processing tasks. In this work, we propose DeepErase, a neural-based preprocessor to erase ink artifacts from text images. We devise a method to programmatically assemble real text images and real artifacts into realistic-looking "dirty" text images, and use them to train an artifact segmentation network in a weakly supervised manner, since pixel-level annotations are automatically obtained during the assembly process. In addition to high segmentation accuracy, we show that our cleansed images achieve a significant boost in recognition accuracy by popular OCR software such as Tesseract 4.0. Finally, we test DeepErase on out-of-distribution datasets (NIST SDB) of scanned IRS tax return forms and achieve double-digit improvements in accuracy. All experiments are performed on both printed and handwritten text.
****

Main-Secondary Network for Defect Segmentation of Textured Surface Images
Yu Xie, Fangrui Zhu, Yanwei Fu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3531-3540
Building an intelligent defect segmentation system for textured images has attracted much increasing attention in both research and industrial communities, due to its significance values in the practical applications of indust

rial inspection and quality control. Previous models learned the classical classifiers for segmentation by designing hand-crafted features. However, defect segmentation of textured surface images poses challenges such as ambiguous shapes and sizes of defects along with varying textures and patterns in the images. Thus, hand-crafted features based segmentation methods can only be applied to particular types of textured images. To this end, it is desirable to learn a general deep learning based representation for the automatic segmentation of defects. Furthermore, it is relatively less study in efficiently extracting the deep features in the frequency domain, which, nevertheless, should be very important to understand the patterns of textured images. In this paper, we propose a novel defect segmentation deep net-work - Main-Secondary Network (MS-Net). Our MS-Net is trained to model both features from the spatial domain and the frequency domain, where wavelet transform is utilized to extract discriminative information from the frequency do-main. Extensive experiments show the effectiveness of our MS-Net.

****

A Novel Inspection System For Variable Data Printing Using Deep Learning

Oren Haik, Oded Perry, Eli Chen, Peter Klammer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3541-3550

We present a novel approach for inspecting variable data prints (VDP) with an ultra-low false alarm rate (0.005%) and potential applicability to other real-world problems. The system is based on a comparison between two images: a reference image and an image captured by low-cost scanners. The comparison task is challenging as low-cost imaging systems create artifacts that may erroneously be classified as true (genuine) defects. To address this challenge we introduce two new fusion methods, for change detection applications, which are both fast and efficient. The first is an early fusion method that combines the two input images into a single pseudo-color image. The second, called Change-Detection Single Shot Detector (CD-SSD) leverages the SSD by fusing features in the middle of the network. We demonstrate the effectiveness of the proposed deep learning-based approach with a large dataset from real-world printing scenarios. Finally, we evaluate our models on a different domain of aerial imagery change detection (AICD). Our best method clearly outperforms the state-of-the-art baseline on this dataset.

****

Print Defect Mapping with Semantic Segmentation

Augusto Valente, Cristina Wada, Deangela Neves, Deangeli Neves, Fabio Perez, Guilherme Megeto, Marcos Cascone, Otavio Gomes, Qian Lin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3551-3559

Efficient automated print defect mapping is valuable to the printing industry since such defects directly influence customer-perceived printer quality and manually mapping them is cost-ineffective. Conventional methods consist of complicated and hand-crafted feature engineering techniques, usually targeting only one type of defect. In this paper, we propose the first end-to-end framework to map print defects at pixel level, adopting an approach based on semantic segmentation. Our framework uses Convolutional Neural Networks, specifically DeepLab-v3+, and achieves promising results in the identification of defects in printed images. We use synthetic training data by simulating two types of print defects and a print-scan effect with image processing and computer graphic techniques. Compared with conventional methods, our framework is versatile, allowing two inference strategies, one being near real-time and providing coarser results, and the other focusing on offline processing with more fine-grained detection. Our model is evaluated on a dataset of real printed images.

****

Low Cost, High Performance Automatic Motorcycle Helmet Violation Detection

Aphinya Chairat, Matthew Dailey, Somphop Limsoonthrakul, Mongkol Ekpanyapong, Dharma Raj KC; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3560-3568

Road fatality rates are very high, especially in developing and middle-income countries. One of the main causes of road fatalities is not using motorcycle helme ts. Active law enforcement may help increase compliance, but ubiquitous enforcement requires many police officers and may cause traffic jams and safety issues. In this paper, we demonstrate the effectiveness of computer vision and machine learning methods to increase helmet compliance through automated helmet violation detection. The system detects riders and passengers not wearing helmets and consists of motorcyclist detection, helmet violation classification, and tracking. The architecture of the system comprises a single GPU server and multiple computational clients that cooperate to complete the task, with communication over HTTP. In a real-world test, the system is able to detect 97% of helmet violations with a 15% false alarm rate. The client-server architecture reduces cost by 20-30% compared to a baseline architecture.

****

Composition-Aware Image Aesthetics Assessment

Dong Liu, Rohit Puri, Nagendra Kamath, Subhabrata Bhattacharya; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3569-3578

Automatic image aesthetics assessment is important for a wide variety of applications such as on-line photo suggestion, photo album management and image retrieval. Previous methods have focused on mapping the holistic image content to a high or low aesthetics rating. However, the composition information of an image characterizes the harmony of its visual elements according to the principles of art, and provides richer information for learning aesthetics. In this work, we propose to model the image composition information as the mutual dependency of its local regions, and design a novel architecture to leverage such information to boost the performance of aesthetics assessment. To achieve this, we densely partition an image into local regions and compute aesthetics-preserving features over the regions to characterize the aesthetics properties of image content. With the feature representation of local regions, we build a region composition graph in which each node denotes one region and any two nodes are connected by an edge weighted by the similarity of the region features. We perform reasoning on this graph via graph convolution, in which the activation of each node is determined by its highly correlated neighbors. Our method naturally uncovers the mutual dependency of local regions in the network training procedure, and achieves the state-of-the-art performance on the benchmark visual aesthetics datasets.

****

Adversarial Defense based on Structure-to-Signal Autoencoders

Sebastian Palacio, Joachim Folz, Jorn Hees, Andreas Dengel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3579-3588

Adversarial attacks have exposed the intricacies of the complex loss surfaces approximated by neural networks. In this paper, we present a defense strategy against gradient-based attacks, on the premise that input gradients need to expose information about the semantic manifold for attacks to be successful. We propose an architecture based on compressive autoencoders (AEs) with a two-stage training scheme, creating not only an architectural bottleneck but also a representational bottleneck. We show that the proposed mechanism yields robust results against a collection of gradient-based attacks under challenging white-box conditions. This defense is attack-agnostic and can, therefore, be used for arbitrary pre-trained models, while not compromising the original performance. These claims are supported by experiments conducted with state-of-the-art image classifiers (ResNet50 and Inception v3), on the full ImageNet validation set. Experiments, including counterfactual analysis, empirically show that the robustness stems from a shift in the distribution of input gradients, which mitigates the effect of tested adversarial attack methods. Gradients propagated through the proposed AEs represent less semantic information and instead point to low-level structural features.

****

Calibrated Domain-Invariant Learning for Highly Generalizable Large Scale Re-Identification

Ye Yuan, Wuyang Chen, Tianlong Chen, Yang Yang, Zhou Ren, Zhangyang Wang, Gang Hua; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3589-3598

Many real-world applications, such as city scale traffic monitoring and control, requires large scale re-identification. However, previous ReID methods often failed to address two limitations in existing ReID benchmarks, i.e. low spatiotemporal coverage and sample imbalance. Notwithstanding their demonstrated success in every single benchmark, they have difficulties in generalizing to unseen environments. As a result, these methods are less applicable in a large scale setting due to poor generalization. In seek for a highly generalizable large-scale ReID method, we present an adversarial domain-invariant feature learning framework (ADIN) that explicitly learns to separate identity-related features from challenging variations, where for the first time "free" annotations in ReID data such as video timestamp and camera index are utilized. We take advantage of the nuisance labels that can be obtained "for free" in ReID data, such as video timestamp and camera index annotations. Furthermore, we find that the imbalance of nuisance classes jeopardizes the adversarial training, and for mitigation we propose a calibrated adversarial loss that is attentive to nuisance distribution. Experiments on existing large-scale person/vehicle ReID datasets demonstrate that ADIN learns more robust and generalizable representations, as evidenced by its outstanding direct transfer performance across datasets, which is a criterion that can better measure the generalizability of large scale Re-ID methods.

****

Two-Grid Preconditioned Solver for Bundle Adjustment

Siddhant Katyan, Shrutimoy Das, Pawan Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3599-3606

We present the design and implementation of Two-Grid Preconditioned Bundle Adjustment (TPBA), a robust and efficient technique for solving the non-linear least squares problem that arises in bundle adjustment. Bundle adjustment (BA) methods for multi-view reconstruction formulate the BA problem as a non-linear least squares problem which is solved by some variant of the traditional Levenberg-Marquardt (LM) algorithm. Most of the computation in LM goes into repeatedly solving the normal equations that arise as a result of linearizing the objective function. To solve these system of equations we use the Generalized Minimal Residual (GMRES) method, which is preconditioned using a deflated algebraic two-grid method. To the best of our knowledge this is the first time that a deflated algebraic two-grid preconditioner has been used along with GMRES, for solving a problem in the computer vision domain. We show that the proposed method is several times faster than the direct method and block Jacobi preconditioned GMRES.

****

Towards a Unified Framework for Visual Compatibility Prediction

Anirudh Singhal, Ayush Chopra, Kumar Ayush, Utkarsh Patel Govind, Balaji Krishnamurthy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3607-3616

Visual compatibility prediction refers to the task of determining if a set of items go well together. Existing techniques for compatibility prediction prioritize sensitivity to type or context in item representations and evaluate using a fill-in-the-blank (FITB) task. We scale the FITB task to stress-test existing methods which highlight the need for a compatibility prediction framework that is sensitive to multiple modalities of item relationships. In this work, we introduce a unified framework for compatibility learning that is jointly conditioned on the type, context, and style. The framework is composed of TC-GAE, a graph-based network that models type & context; SAE, an autoencoder that models style; and a reinforcement-learning based search technique that incorporates these modalities to learn a unified compatibility measure. We conduct experiments on two standard datasets and significantly outperform existing state-of-the-art methods. We also present qualitative analysis and discussions to study the impact of components of the proposed framework.

****

NeurReg: Neural Registration and Its Application to Image Segmentation

Wentao Zhu, Andriy Myronenko, Ziyue Xu, Wenqi Li, Holger Roth, Yufang Huang

, Fausto Milletari, Daguang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3617-3626

Registration is a fundamental task in medical image analysis which can be applied to several tasks including image segmentation, intra-operative tracking, multi-modal image alignment, and motion analysis. Popular registration tools such as ANTs and NiftyReg optimize an objective function for each pair of images from scratch which is time-consuming for large images with complicated deformation. Facilitated by the rapid progress of deep learning, learning-based approaches such as VoxelMorph have been emerging for image registration. These approaches can achieve competitive performance in a fraction of a second on advanced GPUs. In this work, we construct a neural registration framework, called NeurReg, with a hybrid loss of displacement fields and data similarity, which substantially improves the current state-of-the-art of registrations. Within the framework, we simulate various transformations by a registration simulator which generates fixed image and displacement field ground truth for training. Furthermore, we design three segmentation frameworks based on the proposed registration framework: 1) atlas-based segmentation, 2) joint learning of both segmentation and registration tasks, and 3) multi-task learning with atlas-based segmentation as an intermediate feature. Extensive experimental results validate the effectiveness of the proposed NeurReg framework based on various metrics: the endpoint error (EPE) of the predicted displacement field, mean square error (MSE), normalized local cross-correlation (NLCC), mutual information (MI), Dice coefficient, uncertainty estimation, and the interpretability of the segmentation. The proposed NeurReg improves registration accuracy with fast inference speed, which can greatly accelerate related medical image analysis tasks.

****

Enhanced generative adversarial network for 3D brain MRI super-resolution

Jiancong Wang, Yuhua Chen, Yifan Wu, Jianbo Shi, James Gee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3627-3636

Single image super-resolution (SISR) reconstruction for magnetic resonance imaging (MRI) has generated significant interest because of its potential to not only speed up imaging but to improve quantitative processing and analysis of available image data. Generative Adversarial Networks (GAN) have proven to perform well in image recovery tasks. In this work, we followed the GAN framework and developed a generator coupled with discriminator to tackle the task of 3D SISR on T1 brain MRI images. We developed a novel 3D memory-efficient residual-dense block generator (MRDG) that achieves state-of-the-art performance in terms of PSNR (Peak Signal to Noise Ratio), SSIM (Structural Similarity) and NRMSE (Normalized Root Mean Squared Error) metrics. Paired with MRDG, we also designed a pyramid pooling discriminator (PPD) to recover details on different size scales simultaneously. Finally, we introduced model blending, a simple and computational efficient method to balance between image and texture quality in the final output, to the task of SISR on 3D images.

****

HistoNet: Predicting size histograms of object instances

Kishan Sharma, Moritz Gold, Christian Zurbruegg, Laura Leal-Taixe, Jan Dirk Wegner; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3637-3645

We propose to predict histograms of object sizes in crowded scenes directly without any explicit object instance segmentation. What makes this task challenging is the high density of objects (of the same category), which makes instance identification hard. Instead of explicitly segmenting object instances, we show that directly learning histograms of object sizes improves accuracy while using drastically less parameters. This is very useful for application scenarios where explicit, pixel-accurate instance segmentation is not needed, but their lies interest in the overall distribution of instance sizes. Our core applications are in biology, where we estimate the size distribution of soldier fly larvae, and medicine, where we estimate the size distribution of cancer cells as an intermediate step to calculate tumor cellularity score. Given an image with hundreds of smal l object instances, we output the total count and the size histogram. We also provide a new data set for this task, the FlyLarvae data set, which consists of 11,000 larvae instances labeled pixel-wise. Our method results in an overall improvement in the count and size distribution prediction as compared to state-of-the-art instance segmentation method Mask R-CNN.

****

3D semi-supervised learning with uncertainty-aware multi-view co-training

Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, Holger Roth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3646-3655

While making a tremendous impact in various fields, deep neural networks usually require large amounts of labeled data for training which are expensive to collect in many applications, especially in the medical domain. Unlabeled data, on the other hand, is much more abundant. Semi-supervised learning techniques, such as co-training, could provide a powerful tool to leverage unlabeled data. In this paper, we propose a novel framework, uncertainty-aware multi-view co-training (UMCT), to address semi-supervised learning on 3D data, such as volumetric data from medical imaging. In our work, co-training is achieved by exploiting multi-viewpoint consistency of 3D data. We generate different views by rotating or permuting the 3D data and utilize asymmetrical 3D kernels to encourage diversified features in different sub-networks. In addition, we propose an uncertainty-weighted label fusion mechanism to estimate the reliability of each view's prediction with Bayesian deep learning. As one view requires the supervision from other views in co-training, our self-adaptive approach computes a confidence score for the prediction of each unlabeled sample in order to assign a reliable pseudo label. Thus, our approach can take advantage of unlabeled data during training. We show the effectiveness of our proposed semi-supervised method on several public datasets from medical image segmentation tasks (NIH pancreas & LiTS liver tumor dataset). Meanwhile, a fully-supervised method based on our approach achieved state-of-the-art performances on both the LiTS liver tumor segmentation and the Medical Segmentation Decathlon (MSD) challenge, demonstrating the robustness and value of our framework, even when fully supervised training is feasible.

****

IterNet: Retinal Image Segmentation Utilizing Structural Redundancy in Vessel Networks

Liangzhi Li, Manisha Verma, Yuta Nakashima, Hajime Nagahara, Ryo Kawasaki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3656-3665

Retinal vessel segmentation is of great interest for diagnosis of retinal vascular diseases. To further improve the performance of vessel segmentation, we propose IterNet, a new model based on UNet, with the ability to find obscured details of the vessel from the segmented vessel image itself, rather than the raw input image. IterNet consists of multiple iterations of a mini-UNet, which can be 4X deeper than the common UNet. IterNet also adopts the weight-sharing and skip-connection features to facilitate training; therefore, even with such a large architecture, IterNet can still learn from merely 10 20 labeled images, without pre-training or any prior knowledge. IterNet achieves AUCs of 0.9816, 0.9851, and 0.9881 on three mainstream datasets, namely DRIVE, CHASE-DB1, and STARE, respectively, which currently are the best scores in the literature. The source code is available.

****

Star-convex Polyhedra for 3D Object Detection and Segmentation in Microscopy

Martin Weigert, Uwe Schmidt, Robert Haase, Ko Sugawara, Gene Myers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3666-3673

Accurate detection and segmentation of cell nuclei in volumetric (3D) fluorescence microscopy datasets is an important step in many biomedical research projects. Although many automated methods for these tasks exist, they often struggle for images with low signal-to-noise ratios and/or dense packing of nuclei. It was recently shown for 2D microscopy images that these issues can be alleviated by tr aining a neural network to directly predict a suitable shape representation (star-convex polygon) for cell nuclei. In this paper, we adopt and extend this approach to 3D volumes by using star-convex polyhedra to represent cell nuclei and similar shapes. To that end, we overcome the challenges of 1) finding parameter-efficient star-convex polyhedra representations that can faithfully describe cell nuclei shapes, 2) adapting to anisotropic voxel sizes often found in fluorescence microscopy datasets, and 3) efficiently computing intersections between pairs of star-convex polyhedra (required for non-maximum suppression). Although our approach is quite general, since star-convex polyhedra include common shapes like bounding boxes and spheres as special cases, our focus is on accurate detection and segmentation of cell nuclei. Finally, we demonstrate on two challenging datasets that our approach (StarDist-3D) leads to superior results when compared to classical and deep learning based methods.

****

Kornia: an Open Source Differentiable Computer Vision Library for PyTorch

Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, Gary Bradski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2020, pp. 3674-3683

This work presents Kornia -- an open source computer vision library which consists of a set of differentiable routines and modules to solve generic computer vision problems. At its core, the package uses PyTorch as its main backend both for efficiency and to take advantage of the reverse-mode auto-differentiation to define and compute the gradient of complex functions. Inspired by OpenCV, Kornia is composed of a set of modules containing operators that can be inserted inside neural networks to train models to perform image transformations, camera calibration, epipolar geometry, and low level image processing techniques such as filtering and edge detection that operate directly on high dimensional tensor representations. Examples of classical vision problems implemented using our framework are also provided including a benchmark comparing to existing vision libraries.

****