SIZER: A Dataset and Model for Parsing 3D Clothing and Learning Size Sensitive 3D Clothing

Garvita Tiwari1(B), Bharat Lal Bhatnagar1, Tony Tung2, and Gerard Pons-Moll1
1MPI for Informatics, Saarland Informatics Campus, Saarbr¨ ucken, Germany
{gtiwari,bbhatnag,gpons }@mpi-inf.mpg.de
2Facebook Reality Labs, Sausalito, USA
tony.tung@fb.com

Fig. 1. SIZER dataset of people with clothing size variation. (Left): 3D scans of people captured in di■erent clothing styles and sizes.(Right): T-shirt and short pants for sizes small and large, which are registered to a common template.

Abstract. While models of 3D clothing learned from real data exist, no method can predict clothing deformation as a function of garment size. In this paper, we introduce SizerNet to predict 3D clothing conditioned on human body shape and garment size parameters, and ParserNet toinfer garment m eshes and shape under clothing with personal details in a single pass from an input mesh. SizerNet allows to estimate and visualize the dressing e■ect of a garment in various sizes, and ParserNet allowsto edit cl othing of an input mesh directly, removing the need for scan segmentation, which is a challenging problem in itself. To learn these models, we introduce the SIZER dataset of clothing size variation which includes 100 di■erent subjects wearing casual clothing items in various sizes, totaling to approximately 2000 scans. This dataset includes the scans, registrations to the SMPL model, scans segmented in clothing parts, garment category and size labels. Our experiments show better

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58580-8 1) contains supplementary material, which is a vail-able to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12348, pp. 1–18, 2020.https://doi.org/10.1007/978-3-030-58580-8

_
****

LIMP: Learning Latent Shape Representations with Metric Preservation Priors

Luca Cosmo1,2(B), Antonio Norelli1, Oshri Halimi3, Ron Kimmel3, and Emanuele Rodol` a1
1Sapienza University of Rome, Rome, Italy
cosmo@di.uniroma1.it
2University of Lugano, Lugano, Switzerland
3Technion - Israel Institute of Technology, Haifa, Israel

Abstract. In this paper, we advocate the adoption of metric preservation as a powerful prior for learning latent representations of deformable 3D shapes. Key to our construction is the introduction of a geometricdistortion criterion, de■ned directly on the decoded shapes, translating the preservation of the metric on the decoding to the formation of linear paths in the underlying latent space. Our rationale lies in the observa-tion tha t training samples alone are often insu■cient to endow generative models with high ■delity, motivating the need for large training datasets. In contrast, metric preservation provides a rigorous way to control theamount of geometric distortion incurring in the construction of the latent space, leading in turn to synthetic samples of higher quality. We further demonstrate, for the ■rst time, the adoption of di■erentiable intrinsicdistances in the backpropagation of a geodesic loss. Our geometric pri-ors are particularly relevant in the presence of scarce training data, where

learning any meaningful latent structure can be especially challenging.The e■ect iveness and potential of our generative model is showcased in applications of style transfer, content generation, and shape completion.
Keywords: Learning shapes
·Generative model ·Metric distortion
1
****
Unsupervised Sketch to Photo Synthesis
Runtao Liu1, Qian Yu1,2(B), and Stella X. Yu1
1UC Berkeley/ICSI, Berkeley, USA
qianyu@buaa.edu.cn
2Beihang University,
Xueyuan Rd. No. 37, Haidian District, Beijing, China
Abstract. Humans can envision a realistic photo given a free-hand sketch that is not only spatially imprecise and geometrically distorted but also without colors and visual details. We study unsupervised sketchto photo synthesis for the ■rst time, learning from unpaired sketch and photo data where the target photo for a sketch is unknown during train-ing. Existing works only deal with either style di■erence or spatial defor-matio n alone, synthesizing photos from edge-aligned line drawings or transforming shapes within the same modality, e.g., color images. Our insight is to decompose the unsupervised sketch to photo syn-thesis task into two stages of translation: First shape translation from sketches to grayscale photos and then content enrichment from grayscale to color photos. We also incorporate a self-supervised denoising objec-tive and an attention module to handle abstraction and style variations that are speci■c to sketches. Our synthesis is sketch-faithful and photo-realistic, enabling sketch-based image retrieval and automatic sketch gen-eratio n that captures human visual perception beyond the edge map of a photo.
1
****

(a) (b) (c) (d) (a) (b) (c) (d) (e) (f)
Fig. 7. Left: Synthesized results when the edge map is used as the intermediate goal instead of the grayscale photo. (a) Input sketch; (b) Synthesized edge map, (c) Synthe-sized RGB photo using the edge map; (d) Synthesized RGB photo using grayscale (Ours).
Right: Our model can successfully deal with noise sketches, which are not well handled by another attention-based model, UGATIT. For an input sketch (a), our model pro-duce an attention mask (b); (c) and (d) are grayscale images produced by vanilla and our model. (e) and (f) compare ours with the result of UGATIT. (Color ■gure online)
Fig. 8. Comparisons of paired and unpaired training for shape translation. There are four examples. For each example, the 1st one is the input sketch, the 2nd and the 3rdare grayscale images synthesized by Pix2Pix and our model respectively. Note that for each example, although the input sketches are di■erent visually, Pix2Pix produces a similar-looking grayscale image. Our results are more faithful to the sketch.
4.3 Ablation Study
Two-Stage Architecture. Two-stage architecture is the key to the success of our model. This strategy can be easily adapted by other models such as cycle-GAN. Table 2compares the performance of the original cycleGAN and its two-stage version (i.e., cycleGAN is used only for shape translation while the content

enrichment network is the same as ours). The two-stage version outperforms the original cycleGAN by 27.55 (on ShoeV2) and 68.33 (on ChairV2), indicating the signi■cant bene■ts brought by this architectural design.
Edge Map vs. Grayscale as the Intermediate Goal. We choose grayscale as our intermediate goal of translation. As shown in Fig. 1,edge maps could be an alternative since it does not have shape deformation either. We can ■rst translate sketch to an edge map, and then ■ll the edge map with colorful details .
Table 2and Fig. 7show that using the edge map is worse than using the grayscale. Our explanations are: 1)Grayscale images contain more visual details thus can provide more learning signals for training shape translation network; 2)Content enrichment is easier for grayscale as they are closer to color photos than edge maps. The grayscale is also easier to obtain in practice.
Deal with Abstraction and Style Variations. We have discussed the problem encountered during shape translation in Sect. 3.1, and further introduced 1) a self-supervised objective along with noise sketch composition strategies an
****
A Simple Way to Make Neural Networks Robust Against Diverse Image Corruptions
Evgenia Rusak1,2(B), Lukas Schott1,2, Roland S. Zimmermann1,2, Julian Bitterwolf2, Oliver Bringmann1, Matthias Bethge1,2, and Wieland Brendel1,2
1University of T¨ ubingen, T¨ ubingen, Germany
{evgenia.rusak,lukas.schott,roland.zimmermann, oliver.bringmann,matthias.bethge,wieland.brendel }@uni-tuebingen.de
2International Max Planck Research School for Intelligent Systems, T¨ubingen, Germany
julian.bitterwolf@uni-tuebingen.de
Abstract. The human visual system is remarkably robust against a wide range of naturally occurring variations and corruptions like rain or snow. In contrast, the performance of modern image recognitionmodels strongly degrades when evaluated on previously unseen corruptions. Here, we demonstrate that a simple but properly tuned training with additive Gaussian and Speckle noise generalizes surprisingly well tounseen corruptions, easily reaching the state of the art on the corruption benchmark ImageNet-C (with ResNet50) and on MNIST-C. We build on top of these strong baseline results and show that an adversarial train-ing of the recognition model against locally correlated worst-case noise distributions leads to an additional increase in performance. This regularization can be combined with previously proposed defense methodsfor further improvement.
Keywords: Image corruptions ·Robustness ·Generalization · Adversarial training
1
****
SoftPoolNet: Shape Descriptor for Point Cloud Completion and Classi■cation
Yida Wang1(B), David Joseph Tan2, Nassir Navab1, and Federico Tombari1,2
1Technische Universit¨ at M¨unchen, M¨ unchen, Germany
yida.wang@tum.de
2Google Inc., Menlo Park, USA
Abstract. Point clouds are often the default choice for many applications as they exhibit more ■exibility and e■ciency than volumetric data. Nevertheless, their unorganized nature – points are stored in anunordered way – makes them less suited to be processed by deep learning pipelines. In this paper, we propose a method for 3D object completion and classi■cation based on point clouds. We introduce a new way of orga-nizing the extracted features based on their activations, which we name

soft pooling. For the decoder stage, we propose regional convolutions, a
novel operator aimed at maximizing the global activation entropy. Fur-thermore,
inspired by the local re■ning procedure in Point Completion
Network (PCN), we also propose a patch-deforming operation to simu-
late deconvolutional operations for point clouds. This paper proves thatour regi
onal activation can be incorporated in many point cloud architec-
tures like AtlasNet and PCN, leading to better performance for geomet-
ric completion. We evaluate our approach on di■erent 3D tasks such asobject comp
letion and classi■cation, achieving state-of-the-art accuracy.
1
****

Hierarchical Face Aging Through
Disentangled Latent Characteristics

Peipei Li1,3, Huaibo Huang1,3,4,Y i b o H u1, Xiang Wu1,R a nH e1,2,3(B),
and Zhenan Sun1,2,3
1Center for Research on Intelligent Perception and Computing, NLPR, CASIA,
Beijing, China
{peipei.li,huaibo.huang }@cripac.ia.ac.cn, huyibo871079699@gmail.com,
alfredxiangwu@gmail.com, {rhe,znsun }@nlpr.ia.ac.cn
2Center for Excellence in Brain Science and Intelligence Technology,
CAS, Beijing, China
3School of Arti■cial Intelligence, University of Chinese Academy of Sciences,
Beijing, China
4Arti■cial Intelligence Research, CAS, Jiaozhou, Qingdao, China

Abstract. Current age datasets lie in a long-tailed distribution, which
brings di■culties to describe the aging mechanism for the imbalanceages. To alle
viate it, we design a novel facial age prior to guide the aging
mechanism modeling. To explore the age e■ects on facial images, we pro-
pose a Disentangled Adversarial Autoencoder (DAAE) to disentangle thefacial imag
es into three independent factors: age, identity and extraneous
information. To avoid the “wash away” of age and identity information
in face aging process, we propose a hierarchical conditional generator bypassing
 the disentangled identity and age embeddings to the high-level
and low-level layers with class-conditional BatchNorm. Finally, a disen-
tangled adversarial learning mechanism is introduced to boost the imagequality f
or face aging. In this way, when manipulating the age distribu-
tion, DAAE can achieve face aging with arbitrary ages. Further, given an
input face image, the mean value of the learned age posterior distribu-tion can
be treated as an age estimator. These indicate that DAAE can
e■ciently and accurately estimate the age distribution in a disentangling
manner. DAAE is the ■rst attempt to achieve facial age analysis tasks,including
face aging with arbitrary ages, exemplar-based face aging and
age estimation, in a universal framework. The qualitative and quantita-
tive experiments demonstrate the superiority of DAAE on ■ve popular
datasets, including CACD2000, Morph, UTKFace, FG-NET and AgeDB.

Keywords: Facial age analysis
·Variational autoencoder

P. Li, H. Huang and R. He—Equal contribution.

Electronic supplementary material The online version of this chapter (https://
doi.org/10.1007/978-3-030-58580-8 6) contains supplementary material, which is a
vail-
able to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12348, pp. 86–101, 2020.https://doi.or
g/10.1007/978-3-030-58580-8

_
****

Hybrid Models for Open Set Recognition

Hongjie Zhang1,A n gL i2,J i eG u o1, and Yanwen Guo1(B)
1State Key Laboratory for Novel Software Technology, Nanjing University,

Nanjing 210023, China
hjzhang@smail.nju.edu.cn, {guojie,ywguo }@nju.edu.cn
2DeepMind, Mountain View, CA, USA
anglili@google.com
Abstract. Open set recognition requires a classi■er to detect samples not belonging to any of the classes in its training set. Existing methods ■t a probability distribution to the training samples on their embedding space and detect outliers according to this distribution. The embeddingspace is often obtained from a discriminative classi■er. However, such discriminative representation focuses only on known classes, which may not be critical for distinguishing the unknown classes. We argue that the representation space should be jointly learned from the inlier classi■er and the density estimator (served as an outlier detector). We propose theOpenHyb rid framework, which is composed of an encoder to encode the input data into a joint embedding space, a classi■er to classify samples to inlier classes, and a ■ow-based density estimator to detect whether asample b elongs to the unknown category. A typical problem of existing ■ow-based models is that they may assign a higher likelihood to outliers. However, we empirically observe that such an issue does not occur in ourexperime nts when learning a joint representation for discriminative and generative components. Experiments on standard open set benchmarks also reveal that an end-to-end trained OpenHybrid model signi■cantlyoutperforms state-of-the-art methods and ■ow-based baselines.
Keywords: Flow-based model ·Density estimation ·Image classi■cation

****

TopoGAN: A Topology-Aware Generative Adversarial Network

Fan Wang(B), Huidong Liu, Dimitris Samaras, and Chao Chen
Stony Brook University, Stony Brook, NY 11794, USA
{fanwang1,huidliu,samaras }@cs.stonybrook.edu, chao.chen.1@stonybrook.edu
Abstract. Existing generative adversarial networks (GANs) focus on generating realistic images based on CNN-derived image features, but fail to preserve the structural properties of real images. This can be fatal in applications where the underlying structure (e.g.., neurons, vessels,membrane s, and road networks) of the image carries crucial semantic meaning. In this paper, we propose a novel GAN model that learns the topology of real images, i.e., connectedness and loopy-ness. In particular,we in troduce a new loss that bridges the gap between synthetic image distribution and real image distribution in the topological feature space. By optimizing this loss, the generator produces images with the samestructural t opology as real images. We also propose new GAN evaluation metrics that measure the topological realism of the synthetic images. We show in experiments that our method generates synthetic images withrealistic top ology. We also highlight the increased performance that our method brings to downstream tasks such as segmentation.
Keywords: Topology ·Persistent homology ·Generative Adversarial Network

****

Learning to Localize Actions from Moments

Fuchen Long1, Ting Yao2(B), Zhaofan Qiu1, Xinmei Tian1,J i e b oL u o3, and Tao Mei2
1University of Science and Technology of China, Hefei, China
longfc.ustc@gmail.com, zhaofanqiu@gmail.com, xinmei@ustc.edu.cn
2JD AI Research, Beijing, China

tingyao.ustc@gmail.com, tmei@jd.com
3University of Rochester, Rochester, NY, USA
jluo@cs.rochester.edu
Abstract. With the knowledge of action moments (i.e., trimmed video clips that each contains an action instance), humans could routinelylocalize an action temporally in an untrimmed video. Nevertheless, most practical methods still require all training videos to be labeled with temporal annotations (action category and temporal boundary) and developthe models in a fully-supervised manner, despite expensive labeling e■orts and inapplicable to new categories. In this paper, we introduce a new design of transfer learning type to learn action localization for a largeset of action categories, but only on action moments from the categories of interest and temporal annotations of untrimmed videos from a small set of action classes. Speci■cally, we present Action Herald Networks(AherNet) that integrate such design into an one-stage action localization framework. Technically, a weight transfer function is uniquely devised to build the transformation between classi■cation of action moments orforeground video segments and action localization in synthetic contextual moments or untrimmed videos. The context of each moment is learnt through the adversarial mechanism to di■erentiate the generated featuresfrom those of background in untrimmed videos. Extensive experiments are conducted on the learning both across the splits of ActivityNet v1.3 and from THUMOS14 to ActivityNet v1.3. Our AherNet demonstratesthe superiority even comparing to most fully-supervised action localization methods. More remarkably, we train AherNet to localize actions from 600 categories on the leverage of action moments in Kinetics-600and temporal annotations from 200 classes in ActivityNet v1.3.
This work was performed at JD AI Research.
Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58580-8 9) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12348, pp. 137–154, 2020.https://doi.org/10.1007/978-3-030-58580-8

_
****
ForkGAN: Seeing into the Rainy Night
Ziqiang Zheng1,Y a n gW u2(B), Xinran Han3, and Jianbo Shi3
1UISEE Technology (Beijing) Co., Ltd., Beijing, China
zhengziqiang1@gmail.com
2Kyoto University, Kyoto, Japan
wu.yang.8c@kyoto-u.ac.jp
3University of Pennsylvania, Philadelphia, USA
{hxinran,jshi }@seas.upenn.edu
Abstract. We present a ForkGAN for task-agnostic image translation that can boost multiple vision tasks in adverse weather conditions. Three tasks of image localization/retrieval, semantic image segmentation, andobject detection are evaluated. The key challenge is achieving high-quality image translation without any explicit supervision, or task awareness. Our innovation is a fork-shape generator with one encoder and two decoders that disentangles the domain-speci■c and domain-invariant information. We force the cyclic translation between the weather con-ditions to go through a common encoding space, and make sure the encoding features reveal no information about the domains. Experimental results show our algorithm produces state-of-the-art image synthesisresults and boost three vision tasks’ performances in adverse weathers.
Keywords: Light illumination ·Image-to-image translation ·Image synthesis ·Generative adversarial networks

1
****
TCGM: An Information-Theoretic
Framework for Semi-supervised
Multi-modality Learning
Xinwei Sun1, Yilun Xu2, Peng Cao2,Y u q i n gK o n g2(B), Lingjing Hu3,
Shanghang Zhang4(B), and Yizhou Wang2,5
1Microsoft Research-Asia, Beijing, China
xinsun@microsoft.com
2Center on Frontiers of Computing Studies, Advanced Institute of Information
Technology, Department of Computer Science, Peking University, Beijing, China
{xuyilun,caopeng2016,yuqing.kong,Yizhou.Wang }@pku.edu.cn
3Yanjing Medical College, Capital Medical University, Beijing, China
hulj@ccmu.edu.cn
4UC Berkeley, Berkeley, USA
shz@eecs.berkeley.edu
5Deepwise AI Lab, Beijing, China
Abstract. Fusing data from multiple modalities provides more informa-
tion to train machine learning systems. However, it is prohibitively expen-
sive and time-consuming to label each modality with a large amount ofdata, which
leads to a crucial problem of semi-supervised multi-modal
learning. Existing methods su■er from either ine■ective fusion across
modalities or lack of theoretical guarantees under proper assumptions. In
this paper, we propose a novel information-theoretic approach - namely,
Total Correlation GainMaximization (TCGM) – for semi-supervised
multi-modal learning, which is endowed with promising properties: (i) it
can utilize e■ectively the information across di■erent modalities of unla-
beled data points to facilitate training classi■ers of each modality (ii) ithas
theoretical guarantee to identify Bayesian classi■ers, i.e., the ground
truth posteriors of all modalities. Speci■cally, by maximizing TC-induced
loss (namely TC gain) over classi■ers of all modalities, these classi■erscan coo
peratively discover the equivalent class of ground-truth classi-
■ers; and identify the unique ones by leveraging limited percentage of
labeled data. We apply our method to various tasks and achieve state-of-the-art
results, including the news classi■cation (Newsgroup dataset),
emotion recognition (IEMOCAP and MOSI datasets), and disease pre-
diction (Alzheimer’s Disease Neuroimaging Initiative dataset).
X. Sun and Y. Xu—Equal Contribution.
Electronic supplementary material The online version of this chapter (https://
doi.org/10.1007/978-3-030-58580-8 11) contains supplementary material, which is
available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12348, pp. 171–188, 2020.
https://doi.org/10.1007/978-3-030-58580-8 _1
****
ExchNet: A Uni■ed Hashing Network
for Large-Scale Fine-Grained Image
Retrieval
Quan Cui1,3, Qing-Yuan Jiang2, Xiu-Shen Wei3(B), Wu-Jun Li2,
and Osamu Yoshie1
1Graduate School of IPS, Waseda University, Fukuoka, Japan
cui-quan@toki.waseda.jp ,yoshie@waseda.jp
2National Key Laboratory for Novel Software Technology, Department of Computer
Science and Technology, Nanjing University, Nanjing, China
qyjiang24@gmail.com ,liwujun@nju.edu.cn
3Megvii Research Nanjing, Megvii Technology, Nanjing, China
weixs.gm@gmail.com
Abstract. Retrieving content relevant images from a large-scale ■ne-
grained dataset could su■er from intolerably slow query speed and highlyredundan
t storage cost, due to high-dimensional real-valued embeddings

which aim to distinguish subtle visual di■erences of ■ne-grained objects. In this paper, we study the novel ■ne-grained hashing topic to generate compact binary codes for ■ne-grained images, leveraging the search and storage e■ciency of hash learning to alleviate the aforementioned prob-lems. Spe ci■cally, we propose a uni■ed end-to-end trainable network, termed as ExchNet. Based on attention mechanisms and proposed atten-tion constraints, ExchNet can ■rstly obtain both local and global featuresto rep resent object parts and the whole ■ne-grained objects, respectively. Furthermore, to ensure the discriminative ability and semantic meaning's consistency of these part-level features across images, we design a localfeature alignment approach by performing a feature exchanging opera-tion. Later, an alternating learning algorithm is employed to optimize the whole ExchNet and then generate the ■nal binary hash codes. Val-idated by ex tensive experiments, our ExchNet consistently outperforms state-of-the-art generic hashing methods on ■ve ■ne-grained datasets. Moreover, compared with other approximate nearest neighbor methods,ExchNet achie ves the best speed-up and storage reduction, revealing its e■ciency and practicality.

Keywords: Fine-Grained Image Retrieval ·Learning to hash · Feature alignment ·Large-scale image search

Q. Cui, Q.-Y. Jiang—Equal contribution.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58580-8 12) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12348, pp. 189–205, 2020.https://doi.org/10.1007/978-3-030-58580-8 _1

****

TSIT: A Simple and Versatile Framework for Image-to-Image Translation

Liming Jiang1, Changxu Zhang2, Mingyang Huang3, Chunxiao Liu3, Jianping Shi3, and Chen Change Loy1(B)
1Nanyang Technological University, Singapore, Singapore
{liming002,ccloy }@ntu.edu.sg
2University of California, Berkeley, CA, USA
zhangcx@berkeley.edu
3SenseTime Research, Beijing, China
{huangmingyang,liuchunxiao,shijianping }@sensetime.com

Abstract. We introduce a simple and versatile framework for image-to-image translation. We unearth the importance of normalization layers, and provide a carefully designed two-stream generative model with newly proposed feature transformations in a coarse-to-■ne fashion. This allowsmulti-sc ale semantic structure information and style representation to be e■ectively captured and fused by the network, permitting our method to scale to various tasks in both unsupervised and supervised settings. Noadditiona l constraints (e.g., cycle consistency) are needed, contributing to a very clean and simple method. Multi-modal image synthesis with arbitrary style control is made possible. A systematic study compares the proposed method with several state-of-the-art task-speci■c baselines, verifying its e■ectiveness in both perceptual quality and quantitativeevaluation s. GitHub: https://github.com/EndlessSora/TSIT .

1

****

ProxyBNN: Learning Binarized Neural Networks via Proxy Matrices

Xiangyu He1,2, Zitao Mo1, Ke Cheng1,2, Weixiang Xu1,2, Qinghao Hu1, Peisong Wang1, Qingshan Liu4, and Jian Cheng1,2,3(B)
1NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing, China

{xiangyu.he,qinghao.hu,peisong.wang,jcheng }@nlpr.ia.ac.cn
2School of Arti■cial Intelligence, University of Chinese Academy of Sciences, Beijing, China
3Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
4Nanjing University of Information Science and Technology, Nanjing, China
Abstract. Training Binarized Neural Networks (BNNs) is challenging due to the discreteness. In order to e■ciently optimize BNNs through backward propagations, real-valued auxiliary variables are commonly used to accumulate gradient updates. Those auxiliary variables are thendirectly quantized to binary weights in the forward pass, which brings about large quantization errors. In this paper, by introducing an appropriate proxy matrix, we reduce the weights quantization error while cir-cumventing explicit binary regularizations on the full-precision auxiliary variables. Speci■cally, we regard pre-binarization weights as a linear combination of the basis vectors. The matrix composed of basis vectors is referred to as the proxy matrix, and auxiliary variables serve as the coef-■cients of this linear combination. We are the ■rst to empirically identifyand study the e■ectiveness of learning both basis and coe■cients to construct the pre-binarization weights. This new proxy learning contributes to new leading performances on benchmark datasets.
Keywords: Binarized Neural Networks
·Proxy matrix

****

HMOR: Hierarchical Multi-person
Ordinal Relations for Monocular
Multi-person 3D Pose Estimation

Can Wang2, Jiefeng Li1, Wentao Liu2, Chen Qian2, and Cewu Lu1(B)
1Shanghai Jiao Tong University, Shanghai, China
{ljflikit,lucewu }@sjtu.edu.cn
2SenseTime Research, Beijing, China
{wangcan,liuwentao,qianchen }@sensetime.com
Abstract. Remarkable progress has been made in 3D human pose estimation from a monocular RGB camera. However, only a few studies explored 3D multi-person cases. In this paper, we attempt to address thelack of a global perspective of the top-down approaches by introducing a novel form of supervision - Hierarchical Multi-person Ordinal Relations (HMOR) . The HMOR encodes interaction information as the ordinal relations of depths and angles hierarchically, which captures the body-part and joint level semantic and maintains global consistency at the same time. In our approach, an integrated top-down model is designed to lever-age these ordinal relations in the learning process. The integrated model estimates human bounding boxes, human depths, and root-relative 3D poses simultaneously, with a coarse-to-■ne architecture to improve theaccuracy of depth estimation. The proposed method signi■cantly outperforms state-of-the-art methods on publicly available multi-person 3D pose datasets. In addition to superior performance, our method costslower computation complexity and fewer model parameters.
Keywords: 3D human pose
·Ordinal relations ·Integrated model

****

Mask2CAD: 3D Shape Prediction by
Learning to Segment and Retrieve

Weicheng Kuo1,2(B), Anelia Angelova1,2, Tsung-Yi Lin1,2, and Angela Dai3
1Google AI, Mountain View, USA
weicheng@google.com, anelia@google.com, tsungyi@google.com
2Robotics at Google, Munich, Germany
3Technical University of Munich, Munich, Germany

angela.dai@tum.de
Abstract. Object recognition has seen signi■cant progress in the image
domain, with focus primarily on 2D perception. We propose to leverageexisting la
rge-scale datasets of 3D models to understand the underlying
3D structure of objects seen in an image by constructing a CAD-based
representation of the objects and their poses. We present Mask2CAD,which jointly
 detects objects in real-world images and for each detected
object, optimizes for the most similar CAD model and its pose. We con-
struct a joint embedding space between the detected regions of an imagecorrespon
ding to an object and 3D CAD models, enabling retrieval of
CAD models for an input RGB image. This produces a clean, lightweight
representation of the objects in an image; this CAD-based representa-tion ensure
s a valid, e■cient shape representation for applications such
as content creation or interactive scenarios, and makes a step towards
understanding the transformation of real-world imagery to a syntheticdomain. Exp
eriments on real-world images from Pix3D demonstrate the
advantage of our approach in comparison to state of the art. To facilitate
future research, we additionally propose a new image-to-3D baseline onScanNet wh
ich features larger shape diversity, real-world occlusions, and
challenging image views.

****

A Uni■ed Framework of Surrogate Loss
by Refactoring and Interpolation

Lanlan Liu1,2, Mingzhe Wang2, and Jia Deng2(B)
1University of Michigan, Ann Arbor, MI 48105, USA
llanlan@umich.edu
2Princeton University, Princeton, NJ 08544, USA
{mingzhew,jiadeng }@cs.princeton.edu
Abstract. We introduce UniLoss, a uni■ed framework to generate sur-
rogate losses for training deep networks with gradient descent, reducing
the amount of manual design of task-speci■c surrogate losses. Our keyobservation
 is that in many cases, evaluating a model with a perfor-
mance metric on a batch of examples can be refactored into four steps:
from input to real-valued scores, from scores to comparisons of pairs ofscores,
from comparisons to binary variables, and from binary variables
to the ■nal performance metric. Using this refactoring we generate di■er-
entiable approximations for each non-di■erentiable step through inter-polation.
Using UniLoss, we can optimize for di■erent tasks and metrics
using one uni■ed framework, achieving comparable performance com-
pared with task-speci■c losses. We validate the e■ectiveness of UniLosson three
tasks and four datasets. Code is available at https://github.
com/princeton-vl/uniloss .
Keywords: Loss design
·Image classi■cation ·Pose estimation

****

Deep Re■ectance Volumes: Relightable
Reconstructions from Multi-view
Photometric Images

Sai Bi1(B), Zexiang Xu1,2, Kalyan Sunkavalli2, Milo■ sH a ■san2,
Yannick Hold-Geo■roy2, David Kriegman1, and Ravi Ramamoorthi1
1University of California, San Diego, USA
bisai@cs.ucsd.edu
2Adobe Research, San Jose, USA
Abstract. We present a deep learning approach to reconstruct scene
appearance from unstructured images captured under collocated pointlighting. At
the heart of Deep Re■ectance Volumes is a novel volumetric
scene representation consisting of opacity, surface normal and re■ectance
voxel grids. We present a novel physically-based di■erentiable volumeray marchin

g framework to render these scene volumes under arbitrary viewpoint and lighting. This allows us to optimize the scene volumes to minimize the error between their rendered images and the capturedimages. Our method is able to reconstruct real scenes with challeng- ing non-Lambertian re■ectance and complex geometry with occlusions and shadowing. Moreover, it accurately generalizes to novel viewpoints and lighting, including non-collocated lighting, rendering photorealis- tic images that are signi■cantly better than state-of-the-art mesh-based methods. We also show that our learned re■ectance volumes are editable,allowing for modifying the materials of the captured scenes.

Keywords: View synthesis ·Relighting ·Appearance acquisition · Neural rendering

1

****

Memory-Augmented Dense Predictive Coding for Video Representation Learning

Tengda Han(B), Weidi Xie , and Andrew Zisserman
Visual Geometry Group, Department of Engineering Science, University of Oxford, Oxford, UK
{htd,weidi,az }@robots.ox.ac.uk

Abstract. The objective of this paper is self-supervised learning from video, in particular for representations for action recognition. We make the following contributions: (i) We propose a new architecture and learn- ing framework Memory-augmented Dense Predictive Coding (MemDPC) for the task. It is trained with a predictive attention mechanism over the set of compressed memories , such that any future states can always be constructed by a convex combination of the condensed representa-tions, allowi ng to make multiple hypotheses e■ciently. (ii) We inves- tigate visual-only self-supervised video representation learning from RGB frames, or from unsupervised optical ■ow, or both. (iii) We thor-oughly eval uate the quality of the learnt representation on four di■er- ent downstream tasks: action recognition, video retrieval, learning with scarce annotations, and unintentional action classi■cation. In all cases,we demo nstrate state-of-the-art or comparable performance over other approaches with orders of magnitude fewer training data.

1

****

PointMixup: Augmentation for Point Clouds

Yunlu Chen1(B), Vincent Tao Hu1(B), Efstratios Gavves1, Thomas Mensink1,2, Pascal Mettes1, Pengwan Yang1,3, and Cees G. M. Snoek1
1University of Amsterdam, Amsterdam, The Netherlands
{y.chen3,t.hu }@uva.nl
2Google Research, Amsterdam, The Netherlands
3Peking University, Beijing, China

Abstract. This paper introduces data augmentation for point clouds by interpolation between examples. Data augmentation by interpolation has shown to be a simple and e■ective approach in the image domain.Such a mixup is however not directly transferable to point clouds, as we do not have a one-to-one correspondence between the points of two dif- ferent objects. In this paper, we de■ne data augmentation between pointclouds as a shortest path linear interpolation. To that end, we intro- duce PointMixup, an interpolation method that generates new examples through an optimal assignment of the path function between two pointclouds. We p rove that our PointMixup ■nds the shortest path between two point clouds and that the interpolation is assignment invariant and linear. With the de■nition of interpolation, PointMixup allows to intro-duce str ong interpolation-based regularizers such as mixup and manifold

mixup to the point cloud domain. Experimentally, we show the potential of PointMixup for point cloud classi■cation, especially when examplesare scarce, as well as increased robustness to noise and geometric transformations to points. The code for PointMixup and the experimental details are publicly available (Code is available at: https://github.com/yunlu-chen/PointMixup/).
Keywords: Interpolation ·Point cloud classi■cation ·Data augmentation
1
****

Identity-Guided Human Semantic Parsing for Person Re-identi■cation

Kuan Zhu1,2(B), Haiyun Guo1,Z h i w e iL i u1,2, Ming Tang1,3, and Jinqiao Wang1,2
1National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{kuan.zhu,haiyun.guo,zhiwei.liu,tangm,jqwang }@nlpr.ia.ac.cn
2School of Arti■cial Intelligence, University of Chinese Academy of Sciences, Beijing, China
3Shenzhen In■nova Limited, Shenzhen, China

Abstract. Existing alignment-based methods have to employ the pre-trained human parsing models to achieve the pixel-level alignment, andcannot identify the personal belongings (e.g., backpacks and reticule) which are crucial to person re-ID. In this paper, we propose the identity-guided human semantic parsing approach (ISP) to locate both the humanbody parts and personal belongings at pixel-level for aligned person re-ID only with person identity labels. We design the cascaded clustering on feature maps to generate the pseudo-labels of human parts. Speci■cally,for the pixels of all images of a person, we ■rst group them to foreground or background and then group the foreground pixels to human parts. The cluster assignments are subsequently used as pseudo-labels of human parts to supervise the part estimation and ISP iteratively learns the feature maps and groups them. Finally, local features of both humanbody parts and personal belongings are obtained according to the self-learned part estimation, and only features of visible parts are utilized for the retrieval. Extensive experiments on three widely used datasetsvalidate the superiority of ISP over lots of state-of-the-art methods. Our code is available at https://github.com/CASIA-IVA-Lab/ISP-reID .
Keywords: Person re-ID ·Weakly-supervised human parsing · Aligned representation learning
1
****

Learning Gradient Fields for Shape Generation

Ruojin Cai(B), Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge Belongie, Noah Snavely, and Bharath Hariharan
Cornell University, Ithaca, USA
rc844@cornell.edu

Abstract. In this work, we propose a novel technique to generate shapes from point cloud data. A point cloud can be viewed as samples from a distribution of 3D points whose density is concentrated near the surface of the shape. Point cloud generation thus amounts to moving ran-domly sampled points to high-density areas. We generate point clouds by performing stochastic gradient ascent on an unnormalized probability density, thereby moving sampled points toward the high-likelihoodregions. Our model directly predicts the gradient of the log density ■eld and can be trained with a simple objective adapted from score-based generative models. We show that our method can reach state-of-the-artperformance

for point cloud auto-encoding and generation, while also allowing for extraction of a high-quality implicit surface. Code is available at https://github.com/RuojinCai/ShapeGF .

Keywords: 3D generation ·Generative models 1

****

COCO-FUNIT: Few-Shot Unsupervised Image Translation with a Content Conditioned Style Encoder

Kuniaki Saito1,2(B), Kate Saenko1, and Ming-Yu Liu2
1Boston University, Boston, USA
{keisaito,saenko }@bu.edu
2NVIDIA, Santa Clara, USA
mingyul@nvidia.com

Abstract. Unsupervised image-to-image translation intends to learn a mapping of an image in a given domain to an analogous image in adi■erent domain, without explicit supervision of the mapping. Few-shot unsupervised image-to-image translation further attempts to generalize the model to an unseen domain by leveraging example images of theunseen domain provided at inference time. While remarkably successful, existing few-shot image-to-image translation models ■nd it di■cult to preserve the structure of the input image while emulating the appear-ance of the unseen domain, which we refer to as the content loss problem. This is particularly severe when the poses of the objects in the input and example images are very di■erent. To address the issue, we propose a new few-shot image translation model, COCO-FUNIT, which computes the style embedding of the example images conditioned on the inputimage and a new module called the constant style bias. Through extensive experimental validations with comparison to the state-of-the-art, our model shows e■ectiveness in addressing the content loss problem. Code and pretrained models are available at https://nvlabs.github.io/COCO-FUNIT/ .

Keywords: Image-to-image translation ·Generative Adversarial Networks 1

****

Corner Proposal Network for Anchor-Free, Two-Stage Object Detection

Kaiwen Duan1, Lingxi Xie2, Honggang Qi1, Song Bai3, Qingming Huang1,4(B), and Qi Tian2(B)
1University of Chinese Academy of Sciences, Beijing, China
duankaiwen17@mails.ucas.ac.cn, {hgqi,qmhuang }@ucas.ac.cn
2Huawei Inc., Shenzhen, China
198808xc@gmail.com, tian.qi1@huawei.com
3Huazhong University of Science and Technology, Wuhan, China
songbai.site@gmail.com
4Peng Cheng Laboratory, Shenzhen, China

Abstract. The goal of object detection is to determine the class and location of objects in an image. This paper proposes a novel anchor-free, two-stage framework which ■rst extracts a number of object proposals by ■nding potential corner keypoint combinations and then assigns a classlabel to each proposal by a standalone classi■cation stage. We demonstrate that these two stages are e■ective solutions for improving recall and precision, respectively, and they can be integrated into an end-to-endnetwork. Our approach, dubbed Corner Proposal Network (CPN), enjoys the ability to detect objects of various scales and also avoids being confused by a large number of false-positive proposals. On the MS-COCOdataset, CPN achieves an AP of 49 .2% which is competitive among state-

of-the-art object detection methods. CPN also ■ts the scenario of com-
putational e■ciency, which achieves an AP of 41 .6%/39 .7% at 26 .2/43.3
FPS, surpassing most competitors with the same inference speed. Code
is available at https://github.com/Duankaiwen/CPNDet .
Keywords: Object detection
·Anchor-free detector ·Two-stage
detector ·Corner keypoints ·Object proposals
1
****

PhraseClick: Toward Achieving Flexible
Interactive Segmentation by Phrase
and Click
Henghui Ding1(B), Scott Cohen2, Brian Price2, and Xudong Jiang1
1Nanyang Technological University, Singapore, Singapore
{ding0093,exdjiang }@ntu.edu.sg
2Adobe Research, San Jose, USA
{scohen,bprice }@adobe.com
Abstract. Existing interactive object segmentation methods mainly
take spatial interactions such as bounding boxes or clicks as input.However, the
se interactions do not contain information about explicit
attributes of the target-of-interest and thus cannot quickly specify what
the selected object exactly is, especially when there are diverse scalesof candi
date objects or the target-of-interest contains multiple objects.
Therefore, excessive user interactions are often required to reach desir-
able results. On the other hand, in existing approaches attribute infor-mation o
f objects is often not well utilized in interactive segmentation.
We propose to employ phrase expressions as another interaction input to
infer the attributes of target object. In this way, we can 1) leverage spa-tial
clicks to locate the target object and 2) utilize semantic phrases to
qualify the attributes of the target object. Speci■cally, the phrase expres-
sions focus on “what” the target object is and the spatial clicks are incharge o
f “where” the target object is, which together help to accurately
segment the target-of-interest with smaller number of interactions. More-
over, the proposed approach is ■exible in terms of interaction modes and
can e■ciently handle complex scenarios by leveraging the strengths of
each type of input. Our multi-modal phrase+click approach achieves newstate-of-t
he-art performance on interactive segmentation. To the best of
our knowledge, this is the ■rst work to leverage both clicks and phrases
for interactive segmentation.
Keywords: Interactive segmentation
·Click·Phrase ·Flexible ·
Attribute
1
****

Uni■ed Multisensory Perception:
Weakly-Supervised Audio-Visual
Video Parsing
Yapeng Tian1(B), Dingzeyu Li2, and Chenliang Xu1
1University of Rochester, Rochester, USA
{yapengtian,chenliang.xu }@rochester.edu
2Adobe Research, Seattle, USA
dinli@adobe.com
Abstract. In this paper, we introduce a new problem, named audio-
visual video parsing, which aims to parse a video into temporal eventsegments an
d label them as either audible, visible, or both. Such a prob-
lem is essential for a complete understanding of the scene depicted inside
a video. To facilitate exploration, we collect a Look, Listen, and Parse
(LLP) dataset to investigate audio-visual video parsing in a weakly-
supervised manner. This task can be naturally formulated as a Mul-
timodal Multiple Instance Learning (MMIL) problem. Concretely, wepropose a novel

hybrid attention network to explore unimodal and cross-
modal temporal contexts simultaneously. We develop an attentive MMIL
pooling method to adaptively explore useful audio and visual content
from di■erent temporal extent and modalities. Furthermore, we discover
and mitigate modality bias and noisy label issues with an individual-guided lear
ning mechanism and label smoothing technique, respectively.
Experimental results show that the challenging audio-visual video pars-
ing can be achieved even with only video-level weak labels. Our proposedframewor
k can e■ectively leverage unimodal and cross-modal temporal
contexts and alleviate modality bias and noisy labels problems.
Keywords: Audio-visual video parsing
·Weakly-supervised ·LLP
dataset
1
****
Learning Delicate Local Representations
for Multi-person Pose Estimation
Yuanhao Cai1,2, Zhicheng Wang1(B), Zhengxiong Luo1,3, Binyi Yin1,4,
Angang Du1,5, Haoqian Wang2, Xiangyu Zhang1, Xinyu Zhou1, Erjin Zhou1,
and Jian Sun1
1Megvii Inc., Beijing, China
{caiyuanhao,wangzhicheng,zxy,zej,zhangxiangyu,sunjian }@megvii.com
2Tsinghua University, Beijing, China
wanghaoqian@tsinghua.edu.cn
3Chinese Academy of Sciences, Beijing, China
4Beihang University, Beijing, China
5Ocean University of China, Qingdao, China
Abstract. In this paper, we propose a novel method called Residual
Steps Network (RSN). RSN aggregates features with the same spatialsize (Intra-le
vel features) e■ciently to obtain delicate local representa-
tions, which retain rich low-level spatial information and result in precise
keypoint localization. Additionally, we observe the output features con-tribute
di■erently to ■nal performance. To tackle this problem, we propose
an e■cient attention mechanism - Pose Re■ne Machine (PRM) to make a
trade-o■ between local and global representations in output features andfurther
re■ne the keypoint locations. Our approach won the 1st place of
COCO Keypoint Challenge 2019 and achieves state-of-the-art results on
both COCO and MPII benchmarks, without using extra training data andpretrained m
odel. Our single model achieves 78.6 on COCO test-dev, 93.0
on MPII test dataset. Ensembled models achieve 79.2 on COCO test-dev,
77.1 on COCO test-challenge dataset. The source code is publicly avail-able for
further research at https://github.com/caiyuanhao1998/RSN/ .
Keywords: Human pose estimation
·COCO ·MPII·Feature
aggregation ·Attention mechanism
1
****
Learning to Plan with Uncertain
Topological Maps
Edward Beeching1(B), Jilles Dibangoye1, Olivier Simonin1,
and Christian Wolf2
1INRIA Chroma team, CITI Lab. INSA Lyon, Villeurbanne, France
{edward.beeching,jilles.dibangoye,olivier.simonin }@insa-lyon.fr
2Universit´ e de Lyon, INSA-Lyon, LIRIS, CNRS, Lyon, France
christian.wolf@insa-lyon.fr
https://team.inria.fr/chroma/en/
Abstract. We train an agent to navigate in 3D environments using a
hierarchical strategy including a high-level graph based planner and alocal poli
cy. Our main contribution is a data driven learning based app-
roach for planning under uncertainty in topological maps, requiring an

estimate of shortest paths in valued graphs with a probabilistic struc-ture. Whereas classical symbolic algorithms achieve optimal results on noise-less topologies, or optimal results in a probabilistic sense on graphs with probabilistic structure, we aim to show that machine learning canovercome missing information in the graph by taking into account rich high-dimensional node features, for instance visual information available at each location of the map. Compared to purely learned neural whitebox algorithms, we structure our neural model with an inductive bias for dynamic programming based shortest path algorithms, and we show that a particular parameterization of our neural model corresponds tothe Bellman-Ford algorithm. By performing an empirical analysis of our method in simulated photo-realistic 3D environments, we demonstrate that the inclusion of visual features in the learned neural planner out-performs classical symbolic solutions for graph based planning.

Keywords: Visual navigation ·Topological maps ·Graph neural networks

1

****

Neural Design Network: Graphic Layout Generation with Constraints

Hsin-Ying Lee2, Lu Jiang1, Irfan Essa1,4,P h u o n gB .L e1, Haifeng Gong1, Ming-Hsuan Yang1,2,3, and Weilong Yang1(B)
1Google Research, Mountain View, USA
2University of California, Merced, Merced, USA
3Yonsei University, Seoul, South Korea
4Georgia Institute of Technology, Atlanta, USA
weilongyang@google.com

Abstract. Graphic design is essential for visual communication with layouts being fundamental to composing attractive designs. Layout gen-eration di■ers from pixel-level image synthesis and is unique in terms of the requirement of mutual relations among the desired components. We propose a method for design layout generation that can satisfy user-speci■ed constraints. The proposed neural design network (NDN) consists of three modules. The ■rst module predicts a graph with complete relations from a graph with user-speci■ed relations. The second modulegenerates a layout from the predicted graph. Finally, the third module ■ne-tunes the predicted layout. Quantitative and qualitative experiments demonstrate that the generated layouts are visually similar to real design layouts. We also construct real designs based on predicted layouts for a better understanding of the visual quality. Finally, we demonstrate apractical application on layout recommendation.

1

****

Learning Open Set Network with Discriminative Reciprocal Points

Guangyao Chen1, Limeng Qiao1, Yemin Shi1, Peixi Peng1(B),J i aL i2,3, Tiejun Huang1,3, Shiliang Pu4, and Yonghong Tian1,3(B)
1Department of Computer Science and Technology, Peking University, Beijing, China
{gy.chen,qiaolm,pxpeng,yhtian }@pku.edu.cn
2State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University, Beijing, China
3Peng Cheng Laboratory, Shenzhen, China
4Hikvision Research Institute, Hangzhou, China

Abstract. Open set recognition is an emerging research area that aims to simultaneously classify samples from prede■ned classes and identify the rest as ‘unknown’. In this process, one of the key challenges is toreduce the risk of generalizing the inherent characteristics of numerous unknown samples learned from a small amount of known data. In this

paper, we propose a new concept, Reciprocal Point , which is the poten-
tial representation of the extra-class space corresponding to each known
category. The sample can be classi■ed to known or unknown by theotherness with r
eciprocal points. To tackle the open set problem, we
o■er a novel open space risk regularization term. Based on the bounded
space constructed by reciprocal points, the risk of unknown is reducedthrough mu
lti-category interaction. The novel learning framework called
Reciprocal Point Learning (RPL), which can indirectly introduce the
unknown information into the learner with only known classes, so as tolearn more
 compact and discriminative representations. Moreover, we
further construct a new large-scale challenging aircraft dataset for open
set recognition: Aircraft 300 (Air-300). Extensive experiments on mul-
tiple benchmark datasets indicate that our framework is signi■cantly
superior to other existing approaches and achieves state-of-the-art per-
formance on standard open set benchmarks.
1
****

Convolutional Occupancy Networks
Songyou Peng1,2(B), Michael Niemeyer2,3, Lars Mescheder2,4,
Marc Pollefeys1,5, and Andreas Geiger2,3
1ETH Zurich, Zurich, Switzerland
songyou.peng@inf.ethz.ch
2Max Planck Institute for Intelligent Systems, T¨ ubingen, Germany
3University of T¨ ubingen, T¨ ubingen, Germany
4Amazon, T¨ ubingen, Germany
5Microsoft, Zurich, Switzerland
Abstract. Recently, implicit neural representations have gained popu-
larity for learning-based 3D reconstruction. While demonstrating promis-
ing results, most implicit approaches are limited to comparably simple
geometry of single objects and do not scale to more complicated or large-scale s
cenes. The key limiting factor of implicit methods is their simple
fully-connected network architecture which does not allow for integrating
local information in the observations or incorporating inductive biasessuch as t
ranslational equivariance. In this paper, we propose Convolu-
tional Occupancy Networks, a more ■exible implicit representation for
detailed reconstruction of objects and 3D scenes. By combining convo-lutional en
coders with implicit occupancy decoders, our model incorpo-
rates inductive biases, enabling structured reasoning in 3D space. We
investigate the e■ectiveness of the proposed representation by recon-structing c
omplex geometry from noisy point clouds and low-resolution
voxel representations. We empirically ■nd that our method enables the
■ne-grained implicit 3D reconstruction of single objects, scales to largeindoor
scenes, and generalizes well from synthetic to real data.
1
****

Multi-person 3D Pose Estimation
in Crowded Scenes Based
on Multi-view Geometry
He Chen1(B), Pengfei Guo1, Pengfei Li1,G i mH e eL e e2,
and Gregory Chirikjian1,2
1The Johns Hopkins University, Baltimore, USA
{hchen136, pguo4, pli32, gchirik1 }@jhu.edu
2National University of Singapore, Singapore, Singapore
gimhee.lee@comp.nus.edu.sg, mpegre@nus.edu.sg
Abstract. Epipolar constraints are at the core of feature matching and
depth estimation in current multi-person multi-camera 3D human pose
estimation methods. Despite the satisfactory performance of this formu-lation in
 sparser crowd scenes, its e■ectiveness is frequently challenged
under denser crowd circumstances mainly due to two sources of ambi-
guity. The ■rst is the mismatch of human joints resulting from the sim-ple cues

provided by the Euclidean distances between joints and epipo-
lar lines. The second is the lack of robustness from the naive formu-
lation of the problem as a least squares minimization. In this paper,we depart f
rom the multi-person 3D pose estimation formulation, and
instead reformulate it as crowd pose estimation. Our method consists
of two key components: a graph model for fast cross-view matching,
and a maximum a posteriori (MAP) estimator for the reconstruction of
the 3D human poses. We demonstrate the e■ectiveness and superiorityof our propos
ed method on four benchmark datasets. Our code is avail-
able at: https://github.com/HeCraneChen/3D-Crowd-Pose-Estimation-
Based-on-MVG .
Keywords: 3D pose estimation
·Occlusion ·Correspondence problem
1
****
TIDE: A General Toolbox for Identifying
Object Detection Errors
Daniel Bolya(B), Sean Foley, James Hays, and Judy Ho■man
Georgia Institute of Technology, Atlanta, USA
dbolya@gatech.edu
Abstract. We introduce TIDE, a framework and associated toolbox
(https://dbolya.github.io/tide/) for analyzing the sources of error in
object detection and instance segmentation algorithms. Importantly, ourframework
 is applicable across datasets and can be applied directly to
output prediction ■les without required knowledge of the underlying pre-
diction system. Thus, our framework can be used as a drop-in replace-ment for th
e standard mAP computation while providing a comprehen-
sive analysis of each model’s strengths and weaknesses. We segment
errors into six types and, crucially, are the ■rst to introduce a tech-nique for
 measuring the contribution of each error in a way that isolates
its e■ect on overall performance. We show that such a representation
is critical for drawing accurate, comprehensive conclusions through in-depth ana
lysis across 4 datasets and 7 recognition models.
Keywords: Error diagnosis
·Object detection ·Instance segmentation
1
****
PointContrast: Unsupervised Pre-training
for 3D Point Cloud Understanding
Saining Xie1(B), Jiatao Gu1, Demi Guo1, Charles R. Qi1, Leonidas Guibas2,
and Or Litany2
1Facebook AI Research, Menlo Park, USA
xiesaining@gmail.com
2Stanford University, Stanford, USA
Abstract. Arguably one of the top success stories of deep learning is
transfer learning. The ■nding that pre-training a network on a rich source
set (e.g., ImageNet) can help boost performance once ■ne-tuned on a usu-
ally much smaller target set, has been instrumental to many applicationsin langu
age and vision. Yet, very little is known about its usefulness in
3D point cloud understanding. We see this as an opportunity considering
the e■ort required for annotating data in 3D. In this work, we aim atfacilitatin
g research on 3D representation learning. Di■erent from pre-
vious works, we focus on high-level scene understanding tasks. To this
end, we select a suit of diverse datasets and tasks to measure the e■ect ofunsup
ervised pre-training on a large source set of 3D scenes. Our ■ndings
are extremely encouraging: using a uni■ed triplet of architecture, source
dataset, and contrastive loss for pre-training, we achieve improvementover recen
t best results in segmentation and detection across 6 di■erent
benchmarks for indoor and outdoor, real and synthetic datasets – demon-
strating that the learned representation can generalize across domains.Furthermo

re, the improvement was similar to supervised pre-training,
suggesting that future e■orts should favor scaling data collection over
more detailed annotation. We hope these ■ndings will encourage moreresearch on u
nsupervised pretext task design for 3D deep learning.
Keywords: Unsupervised learning
·Point cloud recognition ·
Representation learning ·3D scene understanding
1
****
DSA: More E■cient Budgeted Pruning
via Di■erentiable Sparsity Allocation
Xuefei Ning1, Tianchen Zhao2,W e n s h u oL i1, Peng Lei2,
Yu Wang1(B), and Huazhong Yang2
1Department of Electronic Engineering, Tsinghua University, Beijing, China
foxdoraame@gmail.com, yu-wang@tsinghua.edu.cn
2Department of Electronic Engineering, Beihang University, Beijing, China
ztc16@buaa.edu.cn
Abstract. Budgeted pruning is the problem of pruning under resource
constraints. In budgeted pruning, how to distribute the resources across
layers (i.e., sparsity allocation) is the key problem. Traditional methods
solve it by discretely searching for the layer-wise pruning ratios, whichlacks e
■ciency. In this paper, we propose Di■erentiable Sparsity Allo-
cation (DSA), an e■cient end-to-end budgeted pruning ■ow. Utilizing
an o v e l di■erentiable pruning process , DSA ■nds the layer-wise pruning
ratios with gradient-based optimization . It allocates sparsity in continu-
ous space, which is more e■cient than methods based on discrete evalua-tion and
search. Furthermore, DSA could work in a pruning-from-scratch
manner, whereas traditional budgeted pruning methods are applied to
pre-trained models. Experimental results on CIFAR-10 and ImageNetshow that DSA c
ould achieve superior performance than current itera-
tive budgeted pruning methods, and shorten the time cost of the overall
pruning process by at least 1.5 ×in the meantime.
Keywords: Budgeted pruning
·Structured pruning ·Model
compression
1
****
Circumventing Outliers of AutoAugment
with Knowledge Distillation
Longhui Wei1,2(B), An Xiao1, Lingxi Xie1, Xiaopeng Zhang1, Xin Chen1,3,
and Qi Tian1
1Huawei Inc., Shenzhen, China
weilh2568@gmail.com, {xiaoan1,tian.qi1 }@huawei.com, 198808xc@gmail.com,
zxphistory@gmail.com
2University of Science and Technology of China, Hefei, China
3Tongji University, Shanghai, China
1410452@tongji.edu.cn
Abstract. AutoAugment has been a powerful algorithm that improves
the accuracy of many vision tasks, yet it is sensitive to the operator spaceas w
ell as hyper-parameters, and an improper setting may degenerate
network optimization. This paper delves deep into the working mecha-
nism, and reveals that AutoAugment may remove part of discriminativeinformation
from the training image and so insisting on the ground-truth
label is no longer the best option. To relieve the inaccuracy of supervi-
sion, we make use of knowledge distillation that refers to the output ofa teache
r model to guide network training. Experiments are performed
in standard image classi■cation benchmarks, and demonstrate the e■ec-
tiveness of our approach in suppressing noise of data augmentation andstabilizin
g training. Upon the cooperation of knowledge distillation and
AutoAugment, we claim the new state-of-the-art on ImageNet classi-

■cation with a top-1 accuracy of 85.8%.
Keywords: AutoML
·AutoAugment ·Knowledge distillation
1
****

S2DNet: Learning Image Features for
Accurate Sparse-to-Dense Matching

Hugo Germain1(B), Guillaume Bourmaud2, and Vincent Lepetit1,2
1LIGM, ´Ecole des Ponts, Univ Gustave Ei■el, CNRS, Marne-la-vall´ ee, France
{hugo.germain,vincent.lepetit }@enpc.fr
2Laboratoire IMS, Universit´ e de Bordeaux, Bordeaux, France
guillaume.bourmaud@u-bordeaux.fr

Abstract. Establishing robust and accurate correspondences is a fundamental backbone to many computer vision algorithms. While recent learning-based feature matching methods have shown promising results in providing robust correspondences under challenging conditions, theyare often limited in terms of precision. In this paper, we introduce S2DNet, a novel feature matching pipeline, designed and trained to e■ciently establish both robust and accurate correspondences. By leverag-ing a sparse-to-dense matching paradigm, we cast the correspondence learning problem as a supervised classi■cation task to learn to output highly peaked correspondence maps. We show that S2DNet achievesstate-of-the-art results on the HPatches benchmark, as well as on several long-term visual localization datasets.
Keywords: Feature matching
·Classi■cation ·Visual localization
1
****

RTM3D: Real-Time Monocular
3D Detection from Object Keypoints
for Autonomous Driving

Peixuan Li1,2,3,4,5, Huaici Zhao1,2,4,5(B), Pengfei Liu1,2,3,4,5,
and Feidao Cao1,2,3,4,5
1Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, China
hczhao@sia.cn
2Institutes for Robotics and Intelligent Manufacturing,
Chinese Academy of Sciences, Shenyang, China
3University of Chinese Academy of Sciences, Beijing, China
4Key Laboratory of Opto-Electronic Information Processing,
Chinese Academy of Sciences, Shenyang, China
5Key Lab of Image Understanding and Computer Vision, Shenyang, Liaoning, China

Abstract. In this work, we propose an e■cient and accurate monocular 3D detection framework in single shot. Most successful 3D detectors take the projection constraint from the 3D bounding box to the 2D box as an important component. Four edges of a 2D box provide only fourconstraints a nd the performance deteriorates dramatically with the small error of the 2D detector. Di■erent from these approaches, our method predicts the nine perspective keypoints of a 3D bounding box in image space, and then utilize the geometric relationship of 3D and 2D perspectives to recover the dimension, location, and orientation in 3D space.In this me thod, the properties of the object can be predicted stably even when the estimation of keypoints is very noisy, which enables us to obtain fast detection speed with a small architecture. Training ourmethod onl y uses the 3D properties of the object without any extra annotations, category-speci■c 3D shape priors, or depth maps. Our method is the ■rst real-time system (FPS >24) for monocular image 3D detection while achieves state-of-the-art performance on the KITTI benchmark.
Keywords: Real-time monocular 3D detection
·Autonomous
driving ·Keypoint detection

1
****
Video Object Segmentation with Episodic
Graph Memory Networks
Xiankai Lu1, Wenguan Wang2(B), Martin Danelljan2, Tianfei Zhou1,
Jianbing Shen1, and Luc Van Gool2
1Inception Institute of Arti■cial Intelligence, Abu Dhabi, UAE
carrierlxk@gmail.com
2ETH Zurich, Z¨ urich, Switzerland
wenguanwang.ai@gmail.com
https://github.com/carrierlxk/GraphMemVOS
Abstract. How to make a segmentation model e■ciently adapt to a
speci■c video as well as online target appearance variations is a fun-
damental issue in the ■eld of video object segmentation. In this work, agraph me
mory network is developed to address the novel idea of “learning
to update the segmentation model”. Speci■cally, we exploit an episodic
memory network, organized as a fully connected graph, to store framesas nodes an
d capture cross-frame correlations by edges. Further, learn-
able controllers are embedded to ease memory reading and writing, as
well as maintain a ■xed memory scale. The structured, external mem-ory design en
ables our model to comprehensively mine and quickly store
new knowledge, even with limited visual information, and the di■eren-
tiable memory controllers slowly learn an abstract method for storinguseful repr
esentations in the memory and how to later use these rep-
resentations for prediction, via gradient descent. In addition, the pro-
posed graph memory network yields a neat yet principled framework,which can gene
ralize well to both one-shot and zero-shot video object
segmentation tasks. Extensive experiments on four challenging bench-
mark datasets verify that our graph memory network is able to facilitatethe adap
tation of the segmentation network for case-by-case video object
segmentation.
Keywords: Video segmentation
·Episodic graph memory ·Learn to
update
1
****
Rethinking Bottleneck Structure for
E■cient Mobile Network Design
Daquan Zhou1,2,3(B), Qibin Hou1(B), Yunpeng Chen2, Jiashi Feng1,
and Shuicheng Yan2
1National University of Singapore, Singapore, Singapore
zhoudaquan21@gmail.com, andrewhoux@gmail.com, elefjia@nus.edu.sg
2Yitu Technology, Singapore, Singapore
{yunpeng.chen,shuicheng.yan }@yitu-inc.com
3Institute of Data Science, NUS, Singapore, Singapore
Abstract. The inverted residual block is dominating architecture design
for mobile networks recently. It changes the classic residual bottleneck
by introducing two design rules: learning inverted residuals and usinglinear bot
tlenecks. In this paper, we rethink the necessity of such design
changes and ■nd it may bring risks of information loss and gradient
confusion. We thus propose to ■ip the structure and present a novel bot-tleneck
design, called the sandglass block, that performs identity map-
ping and spatial transformation at higher dimensions and thus allevi-
ates information loss and gradient confusion e■ectively. Extensive exper-iments
demonstrate that, di■erent from the common belief, such bot-
tleneck structure is more bene■cial than the inverted ones for mobile
networks. In ImageNet classi■cation, by simply replacing the invertedresidual bl
ock with our sandglass block without increasing parameters
and computation, the classi■cation accuracy can be improved by more
than 1.7% over MobileNetV2. On Pascal VOC 2007 test set, we observethat there is

also 0.9% mAP improvement in object detection. We fur-
ther verify the e■ectiveness of the sandglass block by adding it into the
search space of neural architecture search method DARTS. With 25%parameter reduc
tion, the classi■cation accuracy is improved by 0.13%
over previous DARTS models. Code can be found at: https://github.
com/zhoudaquan/rethinking
bottleneck design .
Keywords: Sandglass block ·Residual block ·E■cient architecture
design·Image classi■cation
D. Zhou and Q. Hou—Authors contributed equally.D. Zhou—Work done during an inter
nship at Yitu Tech.
Electronic supplementary material The online version of this chapter (https://
doi.org/10.1007/978-3-030-58580-8 40) contains supplementary material, which is
available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12348, pp. 680–697, 2020.https://doi.o
rg/10.1007/978-3-030-58580-8
_4
****

Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks

Je■rey O. Zhang1(B), Alexander Sax1, Amir Zamir3, Leonidas Guibas2,
and Jitendra Malik1
1UC Berkeley, Berkeley, USA
jozhang@berkeley.edu
2Stanford University, Stanford, USA
3Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
http://sidetuning.berkeley.edu
Abstract. When training a neural network for a desired task, one may
prefer to adapt a pre-trained network rather than starting from ran-
domly initialized weights. Adaptation can be useful in cases when train-
ing data is scarce, when a single learner needs to perform multiple tasks,
or when one wishes to encode priors in the network. The most commonlyemployed ap
proaches for network adaptation are ■ne-tuning and using
the pre-trained network as a ■xed feature extractor , among others. In
this paper, we propose a straightforward alternative: side-tuning .S i d e -
tuning adapts a pre-trained network by training a lightweight “side” net-
work that is fused with the (unchanged) pre-trained network via sum-
mation. This simple method works as well as or better than existingsolutions and
it resolves some of the basic issues with ■ne-tuning, ■xed
features, and other common approaches. In particular, side-tuning is less
prone to over■tting, is asymptotically consistent, and does not su■erfrom catast
rophic forgetting in incremental learning. We demonstrate
the performance of side-tuning under a diverse set of scenarios, including
incremental learning (iCIFAR, iTaskonomy), reinforcement learning, imi-tation le
arning (visual navigation in Habitat), NLP question-answering
(SQuAD v2), and single-task transfer learning (Taskonomy), with con-
sistently promising results.
Keywords: Sidetuning
·Finetuning ·Transfer learning ·
Representation learning ·Lifelong learning ·Incremental learning ·
Continual learning
1
****

Towards Part-Aware Monocular 3D Human Pose Estimation: An Architecture Search Approach

Zerui Chen1,3(B), Yan Huang1, Hongyuan Yu1,3,B i nX u e3,K eH a n1,
Yiru Guo5, and Liang Wang1,2,4
1Center for Research on Intelligent Perception and Computing, NLPR, CASIA,

Beijing, China
{zerui.chen,hongyuan.yu,ke.han }@cripac.ia.ac.cn,
{yhuang,wangliang }@nlpr.ia.ac.cn
2Center for Excellence in Brain Science and Intelligence Technology, CAS,
Beijing, China
3School of Arti■cial Intelligence, University of Chinese Academy of Sciences,
Beijing, China
xuebin2018@ia.ac.cn
4Chinese Academy of Sciences, Arti■cial Intelligence Research (CAS-AIR),
Beijing, China
5School of Astronautics, Beihang University, Beijing, China
guoyiru@buaa.edu.cn
Abstract. Even though most existing monocular 3D pose estimation
approaches achieve very competitive results, they ignore the heterogene-ity amon
g human body parts by estimating them with the same network
architecture. To accurately estimate 3D poses of di■erent body parts, we
attempt to build a part-aware 3D pose estimator by searching a set ofnetwork arc
hitectures. Consequently, our model automatically learns to
select a suitable architecture to estimate each body part. Compared to
models built on the commonly used ResNet-50 backbone, it reduces 62%parameters a
nd achieves better performance. With roughly the same
computational complexity as previous models, our approach achieves
state-of-the-art results on both the single-person and multi-person 3Dpose estim
ation benchmarks.
Keywords: 3D pose estimation
·Body parts ·Neural architecture
search

****

REVISE: A Tool for Measuring and
Mitigating Bias in Visual Datasets
Angelina Wang(B), Arvind Narayanan , and Olga Russakovsky
Princeton University, Princeton, USA
angelina.wang@princeton.edu
Abstract. Machine learning models are known to perpetuate and even
amplify the biases present in the data. However, these data biases fre-
quently do not become apparent until after the models are deployed.
To tackle this issue and to enable the preemptive analysis of large-
scale dataset, we present our tool. REVISE (REvealing VIsual biaSEs)is a tool th
at assists in the investigation of a visual dataset, surfacing
potential biases currently along three dimensions: (1) object-based, (2)
gender-based, and (3) geography-based. Object-based biases relate tosize, contex
t, or diversity of object representation. Gender-based metrics
aim to reveal the stereotypical portrayal of people of di■erent genders.
Geography-based analyses consider the representation of di■erent geo-graphic loc
ations. REVISE sheds light on the dataset al.ong these dimen-
sions; the responsibility then lies with the user to consider the cultural
and historical context, and to determine which of the revealed biasesmay be prob
lematic. The tool then further assists the user by suggest-
ing actionable steps that may be taken to mitigate the revealed biases.
Overall, the key aim of our work is to tackle the machine learning biasproblem e
arly in the pipeline. REVISE is available at https://github.
com/princetonvisualai/revise-tool .
Keywords: Dataset bias
·Dataset analysis ·Computer vision fairness

****

Contrastive Learning for Weakly
Supervised Phrase Grounding
Tanmay Gupta1(B), Arash Vahdat3, Gal Chechik2,3, Xiaodong Yang3,

Jan Kautz3, and Derek Hoiem1
1University of Illinois Urbana-Champaign, Champaign, USA
tgupta6@illinois.edu
2Bar Ilan University, Ramat Gan, Israel
3NVIDIA, Santa Clara, USA
Abstract. Phrase grounding, the problem of associating image regions
to caption words, is a crucial component of vision-language tasks. Weshow that p
hrase grounding can be learned by optimizing word-region
attention to maximize a lower bound on mutual information between
images and caption words. Given pairs of images and captions, we max-imize compa
tibility of the attention-weighted regions and the words
in the corresponding caption, compared to non-corresponding pairs of
images and captions. A key idea is to construct e■ective negative cap-tions for
learning through language model guided word substitutions.
Training with our negatives yields a ~10% absolute gain in accuracy
over randomly-sampled negatives from the training data. Our weaklysupervised phr
ase grounding model trained on COCO-Captions shows a
healthy gain of 5 .7% to achieve 76 .7% accuracy on Flickr30K Entities
benchmark. Our code and project material will be available at http://
tanmaygupta.info/info-ground .
Keywords: Mutual information
·InfoNCE ·Grounding ·Attention

****

Collaborative Learning of Gesture
Recognition and 3D Hand Pose
Estimation with Multi-order Feature
Analysis

Siyuan Yang1,2, Jun Liu3(B), Shijian Lu4, Meng Hwa Er2, and Alex C. Kot2
1Rapid-Rich Object Search (ROSE) Lab, Interdisciplinary Graduate Programme,
Nanyang Technological University, Singapore, Singapore
siyuan005@e.ntu.edu.sg
2School of Electrical and Electronic Engineering, Nanyang Technological Universi
ty,
Singapore, Singapore
{emher,eackot }@ntu.edu.sg
3Information Systems Technology and Design Pillar, Singapore University of
Technology and Design, Singapore, Singapore
junliu@sutd.edu.sg
4School of Computer Science and Engineering, Nanyang Technological University,
Singapore, Singapore
shijian.Lu@ntu.edu.sg
Abstract. Gesture recognition and 3D hand pose estimation are two
highly correlated tasks, yet they are often handled separately. In this
paper, we present a novel collaborative learning network for joint gesture
recognition and 3D hand pose estimation. The proposed network exploitsjoint-awar
e features that are crucial for both tasks, with which gesture
recognition and 3D hand pose estimation boost each other to learn highly
discriminative features. In addition, a novel multi-order multi-stream fea-ture
analysis method is introduced which learns posture and multi-order
motion information from the intermediate feature maps of videos e■ec-
tively and e■ciently. Due to the exploitation of joint-aware features incommon,
the proposed technique is capable of learning gesture recogni-
tion and 3D hand pose estimation even when only gesture or pose labels
are available, and this enables weakly supervised network learning withmuch redu
ced data labeling e■orts. Extensive experiments show that our
proposed method achieves superior gesture recognition and 3D hand pose
estimation performance as compared with the state-of-the-art.
Keywords: Gesture recognition
·3D hand pose estimation ·

Multi-order multi-stream feature analysis ·Slow-fast feature analysis · Multi-scale relation
Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58580-8 45) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12348, pp. 769–786, 2020.https://doi.org/10.1007/978-3-030-58580-8
_4
****