Event Camera

发表年限	序号	会议/期刊名	论文题目	论文作者	单位	论文方法是否与图神经/深度/卷积网络/激光雷达/点云相关	论文摘要
2016	1	Proceedings of the IEEE CVPR, 2016, pp. 3755-3763	Panoramic Stereo Videos with a Single Camera	Rajat Aggarwal, Amrisha Vohra, Anoop M. Namboodiri	International Institute of Information Technology-Hyderabad, India.	不相关	We present a practical solution for generating 360 degree stereo panoramic videos using a single camera. Current approaches either use a moving camera that captures multiple images of a scene, which are then stitched together to form the final panorama, or use multiple cameras that are synchronized. A moving camera limits the solution to static scenes, while multi-camera solutions require dedicated calibrated setups. Our approach improves upon the existing solutions in two significant ways: It solves the problem using a single camera, thus minimizing the calibration problem and providing us the ability to convert any digital camera into a panoramic stereo capture device. It captures all the light rays required for stereo panoramas in a single frame using a compact custom designed mirror, thus making the design practical to manufacture and easier to use. We analyze several properties of the design as well as present panoramic stereo and depth
	2	Proceedings of the IEEE CVPR, 2016, pp. 884-892	Simultaneous Optical Flow and Intensity Estimation From an Event Camera	Patrick Bardow, Andrew J. Davison, Stefan Leutenegger	Dyson Robotics Laboratory at Imperial College, Dept. of Computing, Imperial College London, UK	不相关	Event cameras are bio-inspired vision sensors which mimic retinas to measure per-pixel intensity change rather than outputting an actual intensity image. This proposed paradigm shift away from traditional frame cameras offers significant potential advantages: namely avoiding high data rates, dynamic range limitations and motion blur. Unfortunately, however, established computer vision algorithms may not at all be applied directly to event cameras. Methods proposed so far to reconstruct images, estimate optical flow, track a camera and reconstruct a scene come with severe restrictions on the environment or on the motion of the camera, e.g. allowing only rotation. Here, we propose, to the best of our knowledge, the first algorithm to simultaneously recover the motion field and brightness image, while the camera undergoes a generic motion through any scene. Our approach employs minimisation of a cost function that contains the asynchronous event data as well as spatial and temporal regularisation within a sliding window time interval. Our implementation relies on GPU-based optimisation and runs in near real-time. In a series of examples, we demonstrate the successful operation of our framework, including in situations where conventional cameras heavily suffer from dynamic range limitations or motion blur.
	3	Proceedings of the IEEE CVPR, 2016, pp. 1392-1400	Online Multi-Object Tracking via Structural Constraint Event Aggregation	Ju Hong Yoon, Chang-Ryeol Lee, Ming-Hsuan Yang, Kuk-Jin Yoon	Korea Electronics Technology Institute |	不相关	Multi-object tracking (MOT) becomes more challenging when objects of interest have similar appearances. In that case, the motion cues are particularly useful for discriminating multiple objects. However, for online 2D MOT in scenes acquired from moving cameras, observable motion cues are complicated by global camera movements and thus not always smooth or predictable. To deal with such unexpected camera motion for online 2D MOT, a structural motion constraint between objects has been utilized thanks to its robustness to camera motion. In this paper, we propose a new data association method that effectively exploits structural motion constraints in the presence of large camera motion. In addition, to further improve the robustness of data association against mis-detections and clutters, a novel event aggregation approach is developed to integrate structural constraints in assignment costs for online MOT. Experimental results on a large number of datasets demonstrate the effectiveness of the proposed algorithm for online 2D MOT.
	4	Proceedings of the IEEE CVPR,2016, pp. 1772-1780	Mirror Surface Reconstruction Under an Uncalibrated Camera	Kai Han, Kwan-Yee K. Wong, Dirk Schnieders, Miaomiao Liu	The University of Hong Kong, Hong Kong, NICTA and CECS, ANU, Canberra	相关(点云）	This paper addresses the problem of mirror surface reconstruction, and a solution based on observing the reflections of a moving reference plane on the mirror surface is proposed. Unlike previous approaches which require tedious work to calibrate the camera, our method can recover both the camera intrinsics and extrinsics together with the mirror surface from reflections of the reference plane under at least three unknown distinct poses. Our previous work has demonstrated that 3D poses of the reference plane can be registered in a common coordinate system using reflection correspondences established across images. This leads to a bunch of registered 3D lines formed from the reflection correspondences. Given these lines, we first derive an analytical solution to recover the camera projection matrix through estimating the line projection matrix. We then optimize the camera projection matrix by minimizing reprojection errors computed based on a cross-ratio formulation. The mirror surface is finally reconstructed based on the optimized cross-ratio constraint. Experimental results on both synthetic and real data are presented, which
	5	Proceedings of the IEEE CVPR,2016, pp. 1884-1893	They Are Not Equally Reliable: Semantic Event Search Using Differentiated Concept Classifiers	Xiaojun Chang, Yao-Liang Yu, Yi Yang, Eric P. Xing	Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney 2Machine Learning Department, Carnegie Mellon University	不相关	Complex event detection on unconstrained Internet videos has seen much progress in recent years. However, state-of-the-art performance degrades dramatically when the number of positive training exemplars falls short. Since label acquisition is costly, laborious, and time-consuming, there is a real need to consider the much more challenging semantic event search problem, where no example video is given. In this paper, we present a state-of-the-art event search system without any example videos. Relying on the key observation that events (e.g. dog show) are usually compositions of multiple mid-level concepts (e.g. "dog," "theater," and "dog jumping"), we first train a skip-gram model to measure the relevance of each concept with the event of interest. The relevant concept classifiers then cast votes on the test videos but their reliability, due to lack of labeled training videos, has been largely unaddressed. We propose to combine the concept classifiers based on a principled estimate of their accuracy on the unlabeled test videos. A novel warping technique is proposed to improve the performance and an efficient highly-scalable algorithm is provided to quickly solve the resulting optimization. We conduct extensive experiments on the latest TRECVID MEDTest 2014, MEDTest 2013 and CCV datasets, and achieve state-
	6	Proceedings of the IEEE CVPR, 2016, pp. 3025-3033	Camera Calibration From Periodic Motion of a Pedestrian	Shiyao Huang, Xianghua Ying, Jiangpeng Rong, Zeyu Shang, Hongbin Zha	Key Laboratory of Machine Perception (Ministry of Education) School of Electronic Engineering and Computer Science, Center for Information Science Peking University, Beijing 100871, P.R. China	不相关	Camera calibration directly from image sequences of a pedestrian without using any calibration object is a really challenging task and should be well solved in computer vision, especially in visual surveillance. In this paper, we propose a novel camera calibration method based on recovering the three orthogonal vanishing points (TOVPs), just using an image sequence of a pedestrian walking in a straight line, without any assumption of scenes or motions, e.g., control points with known 3D coordinates, parallel or perpendicular lines, non-natural or pre-designed special human motions, as often necessary in previous methods. The traces of shoes of a pedestrian carry more rich and easily detectable metric information than all other body parts in the periodic motion of a pedestrian, but such information is usually overlooked by previous work. In this paper, we employ the images of the toes of the shoes on the ground plane to determine the vanishing point corresponding to the walking direction, and then utilize harmonic conjugate properties in projective geometry to recover the vanishing point corresponding to the perpendicular direction of the walking direction in the horizontal plane and the vanishing point corresponding to the vertical direction. After recovering all of the TOVPs, the intrinsic and extrinsic parameters of the camera can be determined. Experiments on various scenes and viewing angles prove the feasibility and accuracy of the proposed method.
	7	Proceedings of the IEEE CVPR, 2016, pp. 3103-3111	Recognizing Activities of Daily Living With a Wrist-Mounted Camera	Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, Tatsuya Harada	Graduate School of Information Science and Technology, The University of Tokyo	相关(CNN)	We present a novel dataset and a novel algorithm for recognizing activities of daily living (ADL) from a first-person wearable camera. Handled objects are crucially important for egocentric ADL recognition. For specific examination of objects related to users' actions separately from other objects in an environment, many previous works have addressed the detection of handled objects in images captured from head-mounted and chest-mounted cameras. Nevertheless, detecting handled objects is not always easy because they tend to appear small in images. They can be occluded by a user's body. As described herein, we mount a camera on a user's wrist. A wrist-mounted camera can capture handled objects at a large scale, and thus it enables us to skip the object detection process. To compare a wrist-mounted camera and a head-mounted camera, we also developed a novel and publicly available dataset that includes videos and annotations of daily activities captured simultaneously by both cameras. Additionally, we propose a discriminative video representation that retains spatial and temporal information after encoding the frame descriptors
	8	Proceedings of the IEEE CVPR, 2016,pp. 3290-3298	Single Image Camera Calibration With Lenticular Arrays for Augmented Reality	Ian Schillebeeckx, Robert Pless	Washington University in St. Louis 1 Brookings Dr, St. Louis, MO 63130	不相关	We consider the problem of camera pose estimation for a scenario where the camera may have continuous and unknown changes in its focal length. Understanding frame by frame changes in camera focal length is vital to accurately estimating camera pose and vital to accurately render virtual objects in a scene with the correct perspective. However, most approaches to camera calibration require geometric constraints from many frames or the observation of a 3D calibration object --- both of which may not be feasible in augmented reality settings. This paper introduces a calibration objects based on a flat lenticular array that creates a color coded light-field whose observed color changes depending on the angle from which it is viewed. We derive an approach to estimate the focal length of the camera and the relative pose of an object from a single image. We characterize the performance of camera calibration across various focal lengths and camera models, and we demonstrate the advantages of the focal length estimation in rendering a virtual object in a video with constant zooming.

	9	Proceedings of the IEEE CVPR, 2016,pp. 3494-3502	Seeing Behind the Camera: Identifying the Authorship of a Photograph	Christopher Thomas, Adriana Kovashka	Department of Computer Science University of Pittsburgh	相关(CNN)	We introduce the novel problem of identifying the photographer behind a photograph. To explore the feasibility of current computer vision techniques to address this problem, we created a new dataset of over 180,000 images taken by 41 well-known photographers. Using this dataset, we examined the effectiveness of a variety of features (low and high-level, including CNN features) at identifying the photographer. We also trained a new deep convolutional neural network for this task. Our results show that high-level features greatly outperform low-level features. We provide qualitative results using these learned models that give insight into our method's ability to distinguish between photographers, and allow us to draw interesting conclusions about what specific photographers shoot. We also demonstrate two applications of our method.
	10	Proceedings of the IEEE CVPR, 2016,pp. 3698-3706	Amplitude Modulated Video Camera - Light Separation in Dynamic Scenes	Amir Kolaman, Maxim Lvov, Rami Hagege, Hugo Guterman	Electrical and Computer Engineering Department Ben-Gurion University of the Negev	不相关	Controlled light conditions improve considerably the performance of most computer vision algorithms. Dynamic light conditions create varying spatial changes in color and intensity across the scene. These condition, caused by a moving shadow for example, force developers to create algorithms which are robust to such variations. We suggest a computational camera which produces images that are not influenced by environmental variations in light conditions. The key insight is that many years ago, similar difficulties were already solved in radio communication; As a result each channel is immune to interference from other radio channels. Amplitude Modulated (AM) video camera separates the influence of a modulated light from other unknown light sources in the scene; Causing the AM video camera frame to appear the same - independent of the light conditions in which it was taken. We built a prototype of the AM video camera by using off the shelf hardware and tested it. AM video camera was used to demonstrate color constancy, shadow removal and contrast enhancement in real time. We show theoretically and empirically that: 1. the proposed system can produce images with similar noise levels as a standard camera. 2. The images created by such camera are almost completely immune to temporal, spatial and spectral
	11	Proceedings of the IEEE CVPR, 2016,pp. 4049-4057	6D Dynamic Camera Relocalization From Single Reference Image	Wei Feng, Fei-Peng Tian, Qian Zhang, Jizhou Sun	School of Computer Science and Technology, Tianjin University, Tianjin, China	不相关	Dynamic relocalization of 6D camera pose from single reference image is a costly and challenging task that requires delicate hand-eye calibration and precision positioning platform to do 3D mechanical rotation and translation. In this paper, we show that high-quality camera relocalization can be achieved in a much less expensive way. Based on inexpensive platform with unreliable absolute repositioning accuracy (ARA), we propose a hand-eye calibration free strategy to actively relocate camera into the same 6D pose that produces the input reference image, by sequentially correcting 3D relative rotation and translation. We theoretically prove that, by this strategy, both rotational and translational relative pose can be effectively reduced to zero, with bounded unknown hand-eye pose displacement. To conquer 3D rotation and translation ambiguity, this theoretical strategy is further revised to a practical relocalization algorithm with faster convergence rate and more reliability by jointly adjusting 3D relative rotation and translation. Extensive experiments validate the effectiveness and superior accuracy of the proposed approach on laboratory
	12	Proceedings of the IEEE CVPR, 2016,pp. 4095-4103	Camera Calibration From Dynamic Silhouettes Using Motion Barcodes	Gil Ben-Artzi, Yoni Kasten, Shmuel Peleg, Michael Werman	School of Computer Science and Engineering The Hebrew University of Jerusalem, Israel	不相关	Computing the epipolar geometry between cameras with very different viewpoints is often problematic as matching points are hard to find. In these cases, it has been proposed to use information from dynamic objects in the scene for suggesting point and line correspondences. We propose a speed up of about two orders of magnitude, as well as an increase in robustness and accuracy, to methods computing epipolar geometry from dynamic silhouettes based on a new temporal signature, motion barcode for lines. This is a binary temporal sequence for lines, indicating for each frame the existence of at least one foreground pixel on that line. The motion barcodes of two corresponding epipolar lines are very similar so the search for corresponding epipolar lines can be limited to lines having similar barcodes leading to increased speed, accuracy, and robustness in computing the epipolar geometry.
	13	Proceedings of the IEEE CVPR,2016,pp. 4132-4140	Rolling Shutter Camera Relative Pose: Generalized Epipolar Geometry	Yuchao Dai, Hongdong Li, Laurent Kneip	Research School of Engineering, Australian National University ARC Centre of Excellence for Robotic Vision (ACRV)	不相关	The vast majority of modern consumer-grade cameras employ a rolling shutter mechanism. In dynamic geometric computer vision applications such as visual SLAM, the so-called rolling shutter effect therefore needs to be properly taken into account. A dedicated relative pose solver appears to be the first problem to solve, as it is of eminent importance to bootstrap any derivation of multi-view geometry. However, despite its significance, it has received inadequate attention to date. This paper presents a detailed investigation of the geometry of the rolling shutter relative pose problem. We introduce the rolling shutter essential matrix, and establish its link to existing models such as the push-broom cameras, summarized in a clean hierarchy of multi-perspective cameras. The generalization of well-established concepts from epipolar geometry is completed by a definition of the Sampson distance in the rolling shutter case. The work is concluded with a careful investigation of the introduced epipolar geometry for rolling shutter cameras on several dedicated benchmarks.
	14	Proceedings of the IEEE CVPR, 2016,pp. 4688-4696	Learning Online Smooth Predictors for Realtime Camera Planning Using Recurrent Decision Trees	Jianhui Chen, Hoang M. Le, Peter Carr, Yisong Yue, James J. Little	University of British Columbia	不相关	We study the problem of online prediction for realtime camera planning, where the goal is to predict smooth trajectories that correctly track and frame objects of interest (e.g., players in a basketball game). The conventional approach for training predictors does not directly consider temporal consistency, and often produces undesirable jitter. Although post-hoc smoothing (e.g., via a Kalman filter) can mitigate this issue to some degree, it is not ideal due to overly stringent modeling assumptions (e.g., Gaussian noise). We propose a recurrent decision tree framework that can directly incorporate temporal consistency into a data-driven predictor, as well as a learning algorithm that can efficiently learn such temporally smooth models. Our approach does not require any post-processing, making online smooth predictions much easier to generate when the noise model is unknown. We apply our approach to sports broadcasting: given noisy player detections, we learn where the camera should look based on human demonstrations. Our experiments exhibit significant improvements over conventional baselines and showcase the practicality of our approach.
	15	Proceedings of the IEEE CVPR, 2016,pp. 4810-4819	Event-Specific Image Importance	Yufei Wang, Zhe Lin, Xiaohui Shen, Radomir Mech, Gavin Miller, Garrison W. Cottrell	University of California, San Diego	相关(CNN)	When creating a photo album of an event, people typically select a few important images to keep or share. There is some consistency in the process of choosing the important images, and discarding the unimportant ones. Modeling this selection process will assist automatic photo selection and album summarization. In this paper, we show that the selection of important images is consistent among different viewers, and that this selection process is related to the event type of the album. We introduce the concept of event-specific image importance. We collected a new event album dataset with human annotation of the relative image importance with each event album. We also propose a Convolutional Neural Network (CNN) based method to predict the image importance score of a given event album, using a novel rank loss function and a progressive training scheme. Results demonstrate that our method significantly outperforms various baseline
	1	Proceedings of the IEEE CVPR, 2017, pp. 1096-1105	Unified Embedding and Metric Learning for Zero-Exemplar Event Detection	Noureldien Hussein, Efstratios Gavves, Arnold W.M. Smeulders	QUVA Lab, University of Amsterdam	不相关	Event detection in unconstrained videos is conceived as a content-based video retrieval with two modalities: textual and visual. Given a text describing a novel event, the goal is to rank related videos accordingly. This task is zero-exemplar, no video examples are given to the novel event. Related works train a bank of concept detectors on external data sources. These detectors predict confidence scores for test videos, which are ranked and retrieved accordingly. In contrast, we learn a joint space in which the visual and textual representations are embedded. The space casts a novel event as a probability of pre-defined events. Also, it learns to measure the distance between an event and its related videos. Our model is trained end-to-end on publicly available EventNet. When applied to TRECVID Multimedia Event Detection dataset, it outperforms the state-of-the-art by a considerable margin.
	2	Proceedings of the IEEE CVPR, 2017,pp. 1106-1114	A Practical Method for Fully Automatic Intrinsic Camera Calibration Using Directionally Encoded Light	Mahdi Abbaspour Tehrani, Thabo Beeler, Anselm Grundhofer	University of California Irvine	相关	Calibrating the intrinsic properties of a camera is one of the fundamental tasks required for a variety of computer vision and image processing tasks. The precise measurement of focal length, location of the principal point as well as distortion parameters of the lens is crucial, for example, for 3D reconstruction. Although a variety of methods exist to achieve this goal, they are often cumbersome to carry out, require substantial manual interaction, expert knowledge, and a significant operating volume. We propose a novel calibration method based on the usage of directionally encoded light rays for estimating the intrinsic parameters. It enables a fully automatic calibration with a small device mounted close to the front lens element and still enables an accuracy comparable to standard methods even when the lens is focused up to infinity. Our method overcomes the mentioned limitations since it guarantees an accurate calibration without any human intervention while requiring only a limited amount of space. Besides that, the approach also allows to estimate the distance of the focal plane as well as the size of the aperture. We demonstrate the advantages of the proposed method by evaluating several camera/lens configurations using prototypical devices.

3	Proceedings of the IEEE CVPR, 2017, pp. 2253-2262	ER3: A Unified Framework for Event Retrieval, Recognition and Recounting	Zhanning Gao, Gang Hua, Dongqing Zhang, Nebojsa Jojic, Le Wang, Jianru Xue, Nanning Zheng	Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University	相关(CNN)	We develop a unified framework for complex event retrieval, recognition and recounting. The framework is based on a compact video representation that exploits the temporal correlations in image features. Our feature alignment procedure identifies and removes the feature redundancies across frames and outputs an intermediate tensor representation we call video imprint. The video imprint is then fed into a reasoning network, whose attention mechanism parallels that of memory networks used in language modeling. The reasoning network simultaneously recognizes the event category and locates the key pieces of evidence for event recounting. In event retrieval tasks, we show that the compact video representation aggregated from the video imprint achieves significantly better retrieval accuracy compared with existing methods. We also set new state of the art results in event recognition tasks with an additional benefit: The latent structure in our reasoning network highlights the areas of the video imprint and can be directly used for event recounting. As video imprint maps back to locations in the video frames, the network allows not only the identification of key frames but also specific areas inside each frame which are most influential to the decision process.
4	Proceedings of the IEEE CVPR, 2017, pp. 3260-3269	A Multi-View Stereo Benchmark With High-Resolution Images and Multi-Camera Videos	Thomas Schops, Johannes L. Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, Andreas Geiger	Department of Computer Science, ETH Zurich	相关(LiDAR)	Motivated by the limitations of existing multi-view stereo benchmarks, we present a novel dataset for this task. Towards this goal, we recorded a variety of indoor and outdoor scenes using a high-precision laser scanner and captured both high-resolution DSLR imagery as well as synchronized low-resolution stereo videos with varying fields-of-view. To align the images with the laser scans, we propose a robust technique which minimizes photometric errors conditioned on the geometry. In contrast to previous datasets, our benchmark provides novel challenges and covers a diverse set of viewpoints and scene types, ranging from natural scenes to man-made indoor and outdoor environments. Furthermore, we provide data at significantly higher temporal and spatial resolution. Our benchmark is the first to cover the important use case of hand-held mobile devices while also providing high-resolution DSLR camera images. We make our datasets and an online evaluation server available at http://www.eth3d.net.
5	Proceedings of the IEEE CVPR, 2017, pp. 3873-3882	Fast Person Re-Identification via Cross-Camera Semantic Binary Transformation	Jiaxin Chen, Yunhong Wang, Jie Qin, Li Liu, Ling Shao	Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, China	不相关	Numerous methods have been proposed for person re-identification, most of which however neglect the matching efficiency. Recently, several hashing based approaches have been developed to make re-identification more scalable for large-scale gallery sets. Despite their efficiency, these works ignore cross-camera variations, which severely deteriorate the final matching accuracy. To address the above issues, we propose a novel hashing based method for fast person re-identification, namely Cross-camera Semantic Binary Transformation (CSBT). CSBT aims to transform original high-dimensional feature vectors into compact identity-preserving binary codes. To this end, CSBT first employs a subspace projection to mitigate cross-camera variations, by maximizing intra-person similarities and inter-person discrepancies. Subsequently, a binary coding scheme is proposed via seamlessly incorporating both the semantic pairwise relationships and local affinity information. Finally, a joint learning framework is proposed for simultaneous subspace projection learning and binary coding based on discrete alternating optimization. Experimental results on four benchmarks clearly demonstrate the superiority of CSBT over the state-of-the-art methods.
6	Proceedings of the IEEE CVPR, 2017, pp. 4288-4296	Unsupervised Vanishing Point Detection and Camera Calibration From a Single Manhattan Image With Radial Distortion	Michel Antunes, Joao P. Barreto, Djamila Aouada, Bjorn Ottersten	Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg,	不相关	The article concerns the automatic calibration of a camera with radial distortion from a single image. It is known that, under the mild assumption of square pixels and zero skew, lines in the scene project into circles in the image, and three lines suffice to calibrate the camera up to an ambiguity between focal length and radial distortion. The calibration results highly depend on accurate circle estimation, which is hard to accomplish because lines tend to project into short circular arcs. To overcome this problem, we show that, given a short circular arc edge, it is possible to robustly determine a line that goes through the center of the corresponding circle. These lines, henceforth called Lines of Circle Centres (LCCs), are used in a new method that detects sets of parallel lines and estimates the calibration parameters, including the center and amount of distortion, focal length, and camera orientation with respect to the Manhattan frame. Extensive experiments in both semi-synthetic and real images show that our algorithm outperforms state-of-the-art approaches in unsupervised calibration from a single image, while providing more
7	Proceedings of the IEEE CVPR, 2017, pp. 4447-4456	From Local to Global: Edge Profiles to Camera Motion in Blurred Images	Subeesh Vasu, A. N. Rajagopalan	Indian Institute of Technology Madras	不相关	In this work, we investigate the relation between the edge profiles present in a motion blurred image and the underlying camera motion responsible for causing the motion blur. While related works on camera motion estimation (CME) rely on the strong assumption of space-invariant blur, we handle the challenging case of general camera motion. We first show how edge profiles `alone' can be harnessed to perform direct CME from a single observation. While it is routine for conventional methods to jointly estimate the latent image too through alternating minimization, our above scheme is best-suited when such a pursuit is either impractical or inefficacious. For applications that actually favor an alternating minimization strategy, the edge profiles can serve as a valuable cue. We incorporate a suitably derived constraint from edge profiles into an existing blind deblurring framework and demonstrate improved restoration performance. Experiments reveal that this approach yields state-of-the-art results for the blind deblurring problem.
8	Proceedings of the IEEE CVPR,2017, pp. 4457-4466	On-The-Fly Adaptation of Regression Forests for Online Camera Relocalisation	Tommaso Cavallari, Stuart Golodetz, Nicholas A. Lord, Julien Valentin, Luigi Di Stefano, Philip H. S. Torr	Department of Computer Science and Engineering, University of Bologna	相关	Camera relocalisation is an important problem in computer vision, with applications in simultaneous localisation and mapping, virtual/augmented reality and navigation. Common techniques either match the current image against keyframes with known poses coming from a tracker, or establish 2D-to-3D correspondences between keypoints in the current image and points in the scene in order to estimate the camera pose. Recently, regression forests have become a popular alternative to establish such correspondences. They achieve accurate results, but must be trained offline on the target scene, preventing relocalisation in new environments. In this paper, we show how to circumvent this limitation by adapting a pre-trained forest to a new scene on the fly. Our adapted forests achieve relocalisation performance that is on par with that of offline forests, and our approach runs in under 150ms, making it desirable for real-time systems that require online
9	Proceedings of the IEEE CVPR, 2017, pp. 4798-4806	A New Rank Constraint on Multi-View Fundamental Matrices, and Its Application to Camera Location Recovery	Soumyadip Sengupta, Tal Amir, Meirav Galun, Tom Goldstein, David W. Jacobs, Amit Singer, Ronen Basri	University of Maryland, College Park	不相关	Accurate estimation of camera matrices is an important step in structure from motion algorithms. In this paper we introduce a novel rank constraint on collections of fundamental matrices in multi-view settings. We show that in general, with the selection of proper scale factors, a matrix formed by stacking fundamental matrices between pairs of images has rank 6. Moreover, this matrix forms the symmetric part of a rank 3 matrix whose factors relate directly to the corresponding camera matrices. We use this new characterization to produce better estimations of fundamental matrices by optimizing an L1-cost function using Iterative Re-weighted Least Squares and Alternate Direction Method of Multiplier. We further show that this procedure can improve the recovery of camera locations, particularly in multi-view settings in which fewer images are available.
10	Proceedings of the IEEE CVPR, 2017, pp. 4885-4893	Simultaneous Geometric and Radiometric Calibration of a Projector-Camera Pair	Marjan Shahpaski, Luis Ricardo Sapaico, Gaspard Chevassus, Sabine Susstrunk	School of Computer and Communication Sciences, EPFL	不相关	We present a novel method that allows for simultaneous geometric and radiometric calibration of a projector-camera pair. It is simple and does not require specialized hardware. We prewarp and align a specially designed projection pattern onto a printed pattern of different colorimetric properties. After capturing the patterns in several orientations, we perform geometric calibration by estimating the corner locations of the two patterns in different color channels. We perform radiometric calibration of the projector by using the information contained inside the projected squares. We show that our method performs on par with current approaches that all require separate geometric and radiometric calibration, while being more efficient and user friendly.
11	Proceedings of the IEEE CVPR, 2017, pp. 5048-5057	A Wide-Field-Of-View Monocentric Light Field Camera	Donald G. Dansereau, Glenn Schuster, Joseph Ford, Gordon Wetzstein	Stanford University, Department of Electrical Engineering	不相关	Light field (LF) capture and processing are important in an expanding range of computer vision applications, offering rich textural and depth information and simplification of conventionally complex tasks. Although LF cameras are commercially available, no existing device offers wide field-of-view (FOV) imaging. This is due in part to the limitations of fisheye lenses, for which a fundamentally constrained entrance pupil diameter severely limits depth sensitivity. In this work we describe a novel, compact optical design that couples a monocentric lens with multiple sensors using microlens arrays, allowing LF capture with an unprecedented FOV. Leveraging capabilities of the LF representation, we propose a novel method for efficiently coupling the spherical lens and planar sensors, replacing expensive and bulky fiber bundles. We construct a single-sensor LF camera prototype, rotating the sensor relative to a fixed main lens to emulate a wide-FOV multi-sensor scenario. Finally, we describe a processing toolchain, including a convenient spherical LF parameterization, and demonstrate depth estimation and post-capture refocus for indoor and outdoor panoramas with 15 x 15 x 1600 x 200 pixels (72 MPix) and a 138-degree FOV.

2017	12	Proceedings of the IEEE CVPR, 2017, pp. 5087-5096	Image Splicing Detection via Camera Response Function Analysis	Can Chen, Scott McCloskey, Jingyi Yu	University of Delaware Newark, DE, USA	相关	Recent advances on image manipulation techniques have made image forgery detection increasingly more challenging. An important component in such tools is to fake motion and/or defocus blurs through boundary splicing and copy-move operators, to emulate wide aperture and slow shutter effects. In this paper, we present a new technique based on the analysis of the camera response functions (CRF) for efficient and robust splicing and copy-move forgery detection and localization. We first analyze how non-linear CRFs affect edges in terms of the intensity-gradient bivariable histograms. We show distinguishable shape differences on real vs. forged blurs near edges after a splicing operation. Based on our analysis, we introduce a deep-learning framework to detect and localize forged edges. In particular, we show the problem can be transformed to a handwriting recognition problem an resolved by using a convolutional neural network. We generate a large dataset of forged images produced by splicing followed by retouching and comprehensive experiments show our proposed method outperforms the state-of-the-art techniques in accuracy
	13	Proceedings of the IEEE CVPR, 2017, pp. 5097-5105	Illuminant-Camera Communication to Observe Moving Objects Under Strong External Light by Spread Spectrum Modulation	Ryusuke Sagawa, Yutaka Satoh	The National Institute of Advanced Industrial Science and Technology	相关(点云)	Many algorithms of computer vision use light sources to illuminate objects to actively create situation appropriate to extract their characteristics. For example, the shape and reflectance are measured by a projector-camera system, and some human-machine or VR systems use projectors and displays for interaction. As existing active lighting systems usually assume no severe external lights to observe projected lights clearly, it is one of the limitations of active illumination. In this paper, we propose a method of energy-efficient active illumination in an environment with severe external lights. The proposed method extracts the light signals of illuminants by removing external light using spread spectrum modulation. Because an image sequence is needed to observe modulated signals, the proposed method extends signal processing to realize signal detection projected onto moving objects by combining spread spectrum modulation and spatio-temporal filtering. In the experiments, we apply the proposed method to a structured-light system under sunlight, to photometric stereo with external lights, and to insensible image embedding.
	14	Proceedings of the IEEE CVPR, 2017, pp. 5125-5133	Identifying First-Person Camera Wearers in Third-Person Videos	Chenyou Fan, Jangwon Lee, Mingze Xu, Krishna Kumar Singh, Yong Jae Lee, David J. Crandall, Michael S. Ryoo	Indiana University Bloomington	不相关	We consider scenarios in which we wish to perform joint scene understanding, object tracking, activity recognition, and other tasks in scenarios in which multiple people are wearing body-worn cameras while a third-person static camera also captures the scene. To do this, we need to establish person-level correspondences across first- and third-person videos, which is challenging because the camera wearer is not visible from his/her own egocentric video, preventing the use of direct feature matching. In this paper, we propose a new semi-Siamese Convolutional Neural Network architecture to address this novel challenge. We formulate the problem as learning a joint embedding space for first- and third-person videos that considers both spatial- and motion-domain cues. A new triplet loss function is designed to minimize the distance between correct first- and third-person matches while maximizing the distance between incorrect ones. This end-to-end approach performs significantly better than several baselines, in part by learning the first- and third-person features optimized for matching jointly with the distance measure itself.
	15	Proceedings of the IEEE CVPR, 2017, pp. 5391-5399	Event-Based Visual Inertial Odometry	Alex Zihao Zhu, Nikolay Atanasov, Kostas Daniilidis	University of Pennsyvlania	不相关	Event-based cameras provide a new visual sensing model by detecting changes in image intensity asynchronously across all pixels on the camera. By providing these events at extremely high rates (up to 1MHz), they allow for sensing in both high speed and high dynamic range situations where traditional cameras may fail. In this paper, we present the first algorithm to fuse a purely event-based tracking algorithm with an inertial measurement unit, to provide accurate metric tracking of a camera's full 6dof pose. Our algorithm is asynchronous, and provides measurement updates at a rate proportional to the camera velocity. The algorithm selects features in the image plane, and tracks spatiotemporal windows around these features within the event stream. An Extended Kalman Filter with a structureless measurement model then fuses the feature tracks with the output of the IMU. The camera poses from the filter are then used to initialize the next step of the tracker and reject failed tracks. We show that our method successfully tracks camera motion on the Event-Camera Dataset in a number of challenging situations.
	16	Proceedings of the IEEE CVPR, 2017, pp. 5771-5780	Consistent-Aware Deep Learning for Person Re-Identification in a Camera Network	Ji Lin, Liangliang Ren, Jiwen Lu, Jianjiang Feng, Jie Zhou	Department of Automation, Tsinghua University, Beijing, China	相关(CNN)	In this paper, we propose a consistent-aware deep learning (CADL) framework for person re-identification in a camera network. Unlike most existing person re-identification methods which identify whether two body images are from the same person, our approach aims to obtain the maximal correct matches for the whole camera network. Different from recently proposed camera network based re-identification methods which only consider the consistent information in the matching stage to obtain a global optimal association, we exploit such consistent-aware information under a deep learning framework where both feature representation and image matching are automatically learned with certain consistent constraints. Specifically, we reach the global optimal solution and balance the performance between different cameras by optimizing the similarity and association iteratively. Experimental results show that our method obtains significant performance improvement and outperforms the state-of-the-art methods by large margins.
	17	Proceedings of the IEEE CVPR, 2017, pp. 5898-5907	Understanding Traffic Density From Large-Scale Web Camera Data	Shanghang Zhang, Guanhang Wu, Joao P. Costeira, Jose M. F. Moura	Carnegie Mellon University, Pittsburgh, PA, USA	相关(FCN)	Understanding traffic density from large-scale web camera (webcam) videos is a challenging problem because such videos have low spatial and temporal resolution, high occlusion and large perspective. To deeply understand traffic density, we explore both optimization based and deep learning based methods. To avoid individual vehicle detection or tracking, both methods map the dense image feature into vehicle density, one based on rank constrained regression and the other based on fully convolutional networks (FCN). The regression based method learns different weights for different blocks of the image to embed road geometry and significantly reduce the error induced by camera perspective. The FCN based method jointly estimates vehicle density and vehicle count with a residual learning framework to perform end-to-end dense prediction, allowing arbitrary image resolution, and adapting to different vehicle scales and perspectives. We analyze and compare both methods, and get insights from optimization based method to improve deep model. Since existing datasets do not cover all the challenges in our work, we collected and labelled a large-scale traffic video dataset, containing 60 million frames from 212 webcams. Both methods are extensively evaluated and compared on different counting tasks and datasets. FCN based method significantly reduces the mean absolute error (MAE) from 10.99 to 5.31 on the public dataset TRANCOS compared with the state-of-the-art baseline.
	18	Proceedings of the IEEE CVPR, 2017, pp. 5974-5983	Geometric Loss Functions for Camera Pose Regression With Deep Learning	Alex Kendall, Roberto Cipolla	University of Cambridge	相关(CNN)	Deep learning has shown to be effective for robust and real-time monocular image relocalisation. In particular, PoseNet is a deep convolutional neural network which learns to regress the 6-DOF camera pose from a single image. It learns to localize using high level features and is robust to difficult lighting, motion blur and unknown camera intrinsics, where point based SIFT registration fails. However, it was trained using a naive loss function, with hyper-parameters which require expensive tuning. In this paper, we give the problem a more fundamental theoretical treatment. We explore a number of novel loss functions for learning camera pose which are based on geometry and scene reprojection error. Additionally we show how to automatically learn an optimal weighting to simultaneously regress position and orientation. By leveraging geometry, we demonstrate that our technique significantly improves PoseNet's performance across datasets ranging
	19	Proceedings of the IEEE CVPR, 2017, pp. 6684-6692	DSAC - Differentiable RANSAC for Camera Localization	Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, Carsten Rother	TU Dresden	相关(CNN)	RANSAC is an important algorithm in robust optimization and a central building block for many computer vision applications. In recent years, traditionally hand-crafted pipelines have been replaced by deep learning pipelines, which can be trained in an end-to-end fashion. However, RANSAC has so far not been used as part of such deep learning pipelines, because its hypothesis selection procedure is non-differentiable. In this work, we present two different ways to overcome this limitation. The most promising approach is inspired by reinforcement learning, namely to replace the deterministic hypothesis selection by a probabilistic selection for which we can derive the expected loss w.r.t. to all learnable parameters. We call this approach DSAC, the differentiable counterpart of RANSAC. We apply DSAC to the problem of camera localization, where deep learning has so far failed to improve on traditional approaches. We demonstrate that by directly minimizing the expected loss of the output camera poses, robustly estimated by RANSAC, we achieve an increase in accuracy. In the future, any deep learning pipeline can use DSAC

	20	Proceedings of the IEEE CVPR, 2017, pp. 7054-7063	Unsupervised Adaptive Re-Identification in Open World Dynamic Camera Networks	Rameswar Panda, Amran Bhuiyan, Vittorio Murino, Amit K. Roy-Chowdhury	Department of ECE UC Riverside	不相关	Person re-identification is an open and challenging problem in computer vision. Existing approaches have concentrated on either designing the best feature representation or learning optimal matching metrics in a static setting where the number of cameras are fixed in a network. Most approaches have neglected the dynamic and open world nature of the re-identification problem, where a new camera may be temporarily inserted into an existing system to get additional information. To address such a novel and very practical problem, we propose an unsupervised adaptation scheme for re-identification models in a dynamic camera network. First, we formulate a domain perceptive re-identification method based on geodesic flow kernel that can effectively find the best source camera (already installed) to adapt with a newly introduced target camera, without requiring a very expensive training phase. Second, we introduce a transitive inference algorithm for re-identification that can exploit the information from best source camera to improve the accuracy across other camera pairs in a network of multiple cameras. Extensive experiments on four benchmark datasets demonstrate that the proposed approach significantly outperforms the state-of-the-art unsupervised learning based alternatives whilst being extremely efficient to compute.
	21	Proceedings of the IEEE CVPR, 2017, pp. 7243-7252	A Low Power, Fully Event-Based Gesture Recognition System	Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, Jeff Kusnitz, Michael Debole, Steve Esser, Tobi Delbruck, Myron	IBM Research	相关(CNN)	We present the first gesture recognition system implemented end-to-end on event-based hardware, using a TrueNorth neurosynaptic processor to recognize hand gestures in real-time at low power from events streamed live by a Dynamic Vision Sensor (DVS). The biologically inspired DVS transmits data only when a pixel detects a change, unlike traditional frame-based cameras which sample every pixel at a fixed frame rate. This sparse, asynchronous data representation lets event-based cameras operate at much lower power than frame-based cameras. However, much of the energy efficiency is lost if, as in previous work, the event stream is interpreted by conventional synchronous processors. Here, for the first time, we process a live DVS event stream using TrueNorth, a natively event-based processor with 1 million spiking neurons. Configured here as a convolutional neural network (CNN), the TrueNorth chip identifies the onset of a gesture with a latency of 105 ms while consuming less than 200 mW. The CNN achieves 96.5% out-of-sample accuracy on a newly collected DVS dataset (DvsGesture) comprising 11 hand gesture categories from 29 subjects under 3 illumination conditions.
	22	Proceedings of the IEEE CVPR, 2017, pp. 7253-7262	Modeling Sub-Event Dynamics in First-Person Action Recognition	Hasan F. M. Zaki, Faisal Shafait, Ajmal Mian	School of Computer Science and Software Engineering, The University ofWestern Australia	不相关	First-person videos have unique characteristics such as heavy egocentric motion, strong preceding events, salient transitional activities and post-event impacts. Action recognition methods designed for third person videos may not optimally represent actions captured by first-person videos. We propose a method to represent the high level dynamics of sub-events in first-person videos by dynamically pooling features of sub-intervals of time series using a temporal feature pooling function. The sub-event dynamics are then temporally aligned to make a new series. To keep track of how the sub-event dynamics evolve over time, we recursively employ the Fast Fourier Transform on a pyramidal temporal structure. The Fourier coefficients of the segment define the overall video representation. We perform experiments on two existing benchmark first-person video datasets which have been captured in a controlled environment. Addressing this gap, we introduce a new dataset collected from YouTube which has a larger number of classes and a greater diversity of capture conditions thereby more closely depicting real-world challenges in first-person video analysis. We compare our method to state-of-the-art first person and generic video recognition algorithms. Our method consistently outperforms the nearest competitors by 10.3%,
	1	Proceedings of the IEEE CVPR, 2018, pp. 136-144	Hybrid Camera Pose Estimation	Federico Camposeco, Andrea Cohen, Marc Pollefeys, Torsten Sattler	Department of Computer Science, ETH Zurich	不相关	In this paper, we aim to solve the pose estimation problem of calibrated pinhole and generalized cameras w.r.t. a Structure-from-Motion (SfM) model by leveraging both 2D-3D correspon-dences as well as 2D-2D correspondences. Traditional approaches either focus on the use of 2D-3D matches, known as structure-based pose estimation or solely on 2D-2D matches (structure-less pose estimation). Absolute pose approaches are limited in their performance by the quality of the 3D point triangulations as well as the completeness of the 3D model. Relative pose approaches, on the other hand, while being more accurate, also tend to be far more computationally costly and often return dozens of possible solutions. This work aims to bridge the gap between these two paradigms. We propose a new RANSAC-based approach that automatically chooses the best type of solver to use at each iteration in a data-driven way. The solvers chosen by our RANSAC can range from pure structure-based or structure-less solvers, to any possible combination of hybrid solvers (i.e. using both types of matches) in between. A number of these new hybrid minimal solvers are also presented in this paper. Both synthetic and real data experiments show our approach to be as accurate as structure-less approaches, while staying close to the efficiency of structure-based methods.
	2	Proceedings of the IEEE CVPR, 2018, pp. 1731-1740	HATS: Histograms of Averaged Time Surfaces for Robust Event-Based Object Classification	Amos Sironi, Manuele Brambilla, Nicolas Bourdis, Xavier Lagorce, Ryad Benosman	PROPHESEE, Paris, France	不相关	Event-based cameras have recently drawn the attention of the Computer Vision community thanks to their advantages in terms of high temporal resolution, low power consumption and high dynamic range, compared to traditional frame-based cameras. These properties make event-based cameras an ideal choice for autonomous vehicles, robot navigation or UAV vision, among others. However, the accuracy of event-based object classification algorithms, which is of crucial importance for any reliable system working in real-world conditions, is still far behind their frame-based counterparts. Two main reasons for this performance gap are: 1. The lack of effective low-level representations and architectures for event-based object classification and 2. The absence of large real-world event-based datasets. In this paper we address both problems. First, we introduce a novel event-based feature representation together with a new machine learning architecture. Compared to previous approaches, we use local memory units to efficiently leverage past temporal information and build a robust event-based representation. Second, we release the first large real-world event-based dataset for object classification. We compare our method to the state-of-the-art with extensive experiments, showing better classification performance and real-
	3	Proceedings of the IEEE CVPR, 2018, pp. 1983-1992	GeoNet: Unsupervised Learning of Dense Depth, Optical Flow and Camera Pose	Zhichao Yin, Jianping Shi	SenseTime Research	不相关	We propose GeoNet, a jointly unsupervised learning framework for monocular depth, optical flow and ego-motion estimation from videos. The three components are coupled by the nature of 3D scene geometry, jointly learned by our framework in an end-to-end manner. Specifically, geometric relationships are extracted over the predictions of individual modules and then combined as an image reconstruction loss, reasoning about static and dynamic scene parts separately. Furthermore, we propose an adaptive geometric consistency loss to increase robustness towards outliers and non-Lambertian regions, which resolves occlusions and texture ambiguities effectively. Experimentation on the KITTI driving dataset reveals that our scheme achieves state-of-the-art results in all of the three tasks, performing better than previously unsupervised methods and comparably with supervised ones.
	4	Proceedings of the IEEE CVPR, 2018, pp. 2354-2363	A Perceptual Measure for Deep Single Image Camera Calibration	Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, Jean-François Lalonde	Universit´e Laval	相关(CNN)	Most current single image camera calibration methods rely on specific image features or user input, and cannot be applied to natural images captured in uncontrolled settings. We propose inferring directly camera calibration parameters from a single image using a deep convolutional neural network. This network is trained using automatically generated samples from a large-scale panorama dataset, and considerably outperforms other methods, including recent deep learning-based approaches, in terms of standard L2 error. However, we argue that in many cases it is more important to consider how humans perceive errors in camera estimation. To this end, we conduct a large-scale human perception study where we ask users to judge the realism of 3D objects composited with and without ground truth camera calibration. Based on this study, we develop a new perceptual measure for camera calibration, and demonstrate that our deep calibration network outperforms other methods on this measure. Finally, we demonstrate the use of our calibration network for a number of applications including virtual object insertion, image retrieval and compositing.

2018	5	Proceedings of the IEEE CVPR, 2018, pp. 2616-2625	Geometry-Aware Learning of Maps for Camera Localization	Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, Jan Kautz	Georgia Institute of Technology	相关(DNN)	Maps are a key component in image-based camera localization and visual SLAM systems: they are used to establish geometric constraints between images, correct drift in relative pose estimation, and relocalize cameras after lost tracking. The exact definitions of maps, however, are often application-specific and hand-crafted for different scenarios (e.g. 3D landmarks, lines, planes, bags of visual words). We propose to represent maps as a deep neural net called MapNet, which enables learning a data-driven map representation. Unlike prior work on learning maps, MapNet exploits cheap and ubiquitous sensory inputs like visual odometry and GPS in addition to images and fuses them together for camera localization. Geometric constraints expressed by these inputs, which have traditionally been used in bundle adjustment or pose-graph optimization, are formulated as loss terms in MapNet training and also used during inference. In addition to directly improving localization accuracy, this allows us to update the MapNet (i.e., maps) in a self-supervised manner using additional unlabeled video sequences from the scene. We also propose a novel parameterization for camera rotation which is better suited for deep-learning based camera pose regression. Experimental results on both the indoor 7-Scenes and the outdoor Oxford RobotCar datasets show significant improvement over prior work. The MapNet project webpage is https://goo.gl/mRB3Au.
	6	Proceedings of the IEEE CVPR, 2018, pp. 2868-2876	Estimation of Camera Locations in Highly Corrupted Scenarios: All About That Base, No Shape Trouble	Yunpeng Shi, Gilad Lerman	University ofMinnesota	不相关	We propose a strategy for improving camera location estimation in structure from motion. Our setting assumes highly corrupted pairwise directions (i.e., normalized relative location vectors), so there is a clear room for improving current state-of-the-art solutions for this problem. Our strategy identifies severely corrupted pairwise directions by using a geometric consistency condition. It then selects a cleaner set of pairwise directions as a preprocessing step for common solvers. We theoretically guarantee the successful performance of a basic version of our strategy under a synthetic corruption model. Numerical results on artificial and real data demonstrate the significant improvement obtained by our strategy.
	7	Proceedings of the IEEE CVPR, 2018, pp. 2984-2992	Camera Pose Estimation With Unknown Principal Point	Viktor Larsson, Zuzana Kukelova, Yinqiang Zheng	Lund University Lund, Sweden	不相关	To estimate the 6-DoF extrinsic pose of a pinhole camera with partially unknown intrinsic parameters is a critical sub-problem in structure-from-motion and camera localization. In most of existing camera pose estimation solvers, the principal point is assumed to be in the image center. Unfortunately, this assumption is not always true, especially for asymmetrically cropped images. In this paper, we develop the first exactly minimal solver for the case of unknown principal point and focal length by using four and a half point correspondences (P4.5Pfuv). We also present an extremely fast solver for the case of unknown aspect ratio (P5Pfuva). The new solvers outperform the previous state-of-the-art in terms of stability and speed. Finally, we explore the extremely challenging case of both unknown principal point and radial distortion, and develop the first practical non-minimal solver by using seven point correspondences (P7Pfruv). Experimental results on both simulated data and real Internet images demonstrate the usefulness of our new solvers.
	8	Proceedings of the IEEE CVPR, 2018, pp. 3867-3876	A Unifying Contrast Maximization Framework for Event Cameras, With Applications to Motion, Depth, and Optical Flow Estimation	Guillermo Gallego, Henri Rebecq, Davide Scaramuzza	Dept. of Informatics and Neuroinformatics, University of Zurich and ETH Zurich	相关	We present a unifying framework to solve several computer vision problems with event cameras: motion, depth and optical flow estimation. The main idea of our framework is to find the point trajectories on the image plane that are best aligned with the event data by maximizing an objective function: the contrast of an image of warped events. Our method implicitly handles data association between the events, and therefore, does not rely on additional appearance information about the scene. In addition to accurately recovering the motion parameters of the problem, our framework produces motion-corrected edge-like images with high dynamic range that can be used for further scene analysis. The proposed method is not only simple, but more importantly, it is, to the best of our knowledge, the first method that can be successfully applied to such a diverse set of important vision tasks with event cameras.
	9	Proceedings of the IEEE CVPR, 2018, pp. 4588-4596	Solving the Perspective-2-Point Problem for Flying-Camera Photo Composition	Ziquan Lan, David Hsu, Gim Hee Lee	NUS Graduate School for Integrative Sciences and Engineering	不相关	Drone-mounted flying cameras will revolutionize photo-taking. The user, instead of holding a camera in hand and manually searching for a viewpoint, will interact directly with image contents in the viewfinder through simple gestures, and the flying camera will achieve the desired viewpoint through the autonomous flying capability of the drone. This work studies the underlying viewpoint search problem for composing a photo with two objects of interest, a common situation in photo-taking. We model it as a Perspective-2-Point (P2P) problem, which is under-constrained to determine the six degrees-of-freedom camera pose uniquely. By incorporating the user's composition requirements and minimizing the camera's flying distance, we form a constrained nonlinear optimization problem and solve it in closed form. Experiments on synthetic data sets and on a real flying camera system indicate promising results.
	10	Proceedings of the IEEE CVPR, 2018, pp. 4654-4662	Learning Less Is More - 6D Camera Localization via 3D Surface Regression	Eric Brachmann, Carsten Rother	Visual Learning Lab Heidelberg University (HCI/IWR)	相关(CNN)	Popular research areas like autonomous driving and augmented reality have renewed the interest in image-based camera localization. In this work, we address the task of predicting the 6D camera pose from a single RGB image in a given 3D environment. With the advent of neural networks, previous works have either learned the entire camera localization process, or multiple components of a camera localization pipeline. Our key contribution is to demonstrate and explain that learning a single component of this pipeline is sufficient. This component is a fully convolutional neural network for densely regressing so-called scene coordinates, defining the correspondence between the input image and the 3D scene space. The neural network is prepended to a new end-to-end trainable pipeline. Our system is efficient, highly accurate, robust in training, and exhibits outstanding generalization capabilities. It exceeds state-of-the-art consistently on indoor and outdoor datasets. Interestingly, our approach surpasses existing techniques even without utilizing a 3D model of the scene during training, since the network is able to discover 3D scene geometry
	11	Proceedings of the IEEE CVPR, 2018, pp. 4824-4833	Rolling Shutter and Radial Distortion Are Features for High Frame Rate Multi-Camera Tracking	Akash Bapat, True Price, Jan-Michael Frahm	Department of Computer Science, The University of North Carolina at Chapel Hill	相关	Traditionally, camera-based tracking approaches have treated rolling shutter and radial distortion as imaging artifacts that have to be overcome and corrected for in order to apply standard camera models and scene reconstruction methods. In this paper, we introduce a novel multi-camera tracking approach that for the first time jointly leverages the information introduced by rolling shutter and radial distortion as a feature to achieve superior performance with respect to high-frequency camera pose estimation. In particular, our system is capable of attaining high tracking rates that were previously unachievable. Our approach explicitly leverages rolling shutter capture and radial distortion to process individual rows, rather than entire image frames, for accurate camera motion estimation. We estimate a per-row 6 DoF pose of a rolling shutter camera by tracking multiple points on a radially distorted row whose rays span a curved surface in 3D space. Although tracking systems for rolling shutter cameras exist, we are the first to leverage radial distortion to measure a per-row pose -- enabling us to use less than half the number of cameras required by the previous state of the art. We validate our system on both synthetic and real imagery.
	12	Proceedings of the IEEE CVPR, 2018, pp. 5030-5039	WILDTRACK: A Multi-Camera HD Dataset for Dense Unscripted Pedestrian Detection	Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, François Fleuret	Machine Learning group, Idiap Research Institute & Ecole Polytechnique Federale de Lausanne	不相关	People detection methods are highly sensitive to occlusions between pedestrians, which are extremely frequent in many situations where cameras have to be mounted at a limited height. The reduction of camera prices allows for the generalization of static multi-camera set-ups. Using joint visual information from multiple synchronized cameras gives the opportunity to improve detection performance. In this paper, we present a new large-scale and high-resolution dataset. It has been captured with seven static cameras in a public open area, and unscripted dense groups of pedestrians standing and walking. Together with the camera frames, we provide an accurate joint (extrinsic and intrinsic) calibration, as well as 7 series of 400 annotated frames for detection at a rate of 2 frames per second. This results in over 40,000 bounding boxes delimiting every person present in the area of interest, for a total of more than 300 individuals. We provide a series of benchmark results using baseline algorithms published over the recent months for multi-view detection with deep neural networks, and trajectory estimation using a non-Markovian model.
	13	Proceedings of the IEEE CVPR, 2018, pp. 5157-5166	Camera Style Adaptation for Person Re-Identification	Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, Yi Yang	Cognitive Science Department, Xiamen University, China	相关(CNN)	Being a cross-camera retrieval task, person re-identification suffers from image style variations caused by different cameras. The art implicitly addresses this problem by learning a camera-invariant descriptor subspace. In this paper, we explicitly consider this challenge by introducing camera style (CamStyle) adaptation. CamStyle can serve as a data augmentation approach that smooths the camera style disparities. Specifically, with CycleGAN, labeled training images can be style-transferred to each camera, and, along with the original training samples, form the augmented training set. This method, while increasing data diversity against over-fitting, also incurs a considerable level of noise. In the effort to alleviate the impact of noise, the label smooth regularization (LSR) is adopted. The vanilla version of our method (without LSR) performs reasonably well on few-camera systems in which over-fitting often occurs. With LSR, we demonstrate consistent improvement in all systems regardless of the extent of over-fitting. We also report competitive accuracy compared with the state of the art.Code is available

	14	Proceedings of the IEEE CVPR, 2018, pp. 5419-5427	Event-Based Vision Meets Deep Learning on Steering Prediction for Self-Driving Cars	Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, Davide Scaramuzza	Grupo de Tratamiento de Im´agenes, Universidad Polit´ecnica de Madrid, Spain	相关	Event cameras are bio-inspired vision sensors that naturally capture the dynamics of a scene, filtering out redundant information. This paper presents a deep neural network approach that unlocks the potential of event cameras on a challenging motion-estimation task: prediction of a vehicle's steering angle. To make the best out of this sensor–algorithm combination, we adapt state-of-the-art convolutional architectures to the output of event sensors and extensively evaluate the performance of our approach on a publicly available large scale event-camera dataset (≈1000 km). We present qualitative and quantitative explanations of why event cameras allow robust steering prediction even in cases where traditional cameras fail, e.g. challenging illumination conditions and fast motion. Finally, we demonstrate the advantages of leveraging transfer learning from traditional to event-based vision, and show that our approach outperforms state-of-the-art algorithms
	15	Proceedings of the IEEE CVPR, 2018, pp. 6036-6046	Features for Multi-Target Multi-Camera Tracking and Re-Identification	Ergys Ristani, Carlo Tomasi	Duke University Durham, NC, USA	相关	Multi-Target Multi-Camera Tracking (MTMCT) tracks many people through video taken from several cameras. Person Re-Identification (Re-ID) retrieves from a gallery images of people similar to a person query image. We learn good features for both MTMCT and Re-ID with a convolutional neural network. Our contributions include an adaptive weighted triplet loss for training and a new technique for hard-identity mining. Our method outperforms the state of the art both on the DukeMTMC benchmarks for tracking, and on the Market-1501 and DukeMTMC-ReID benchmarks for Re-ID. We examine the correlation between good Re-ID and good MTMCT scores, and perform ablation studies to elucidate the contributions of the main components of our system. Code is available.
	16	Proceedings of the IEEE CVPR, 2018, pp. 7532-7542	A Low Power, High Throughput, Fully Event-Based Stereo System	Alexander Andreopoulos, Hirak J. Kashyap, Tapan K. Nayak, Arnon Amir, Myron D. Flickner	IBM Research	不相关	We introduce a stereo correspondence system implemented fully on event-based digital hardware, using a fully graph-based non von-Neumann computation model, where no frames, arrays, or any other such data-structures are used. This is the first time that an end-to-end stereo pipeline from image acquisition and rectification, multi-scale spatio-temporal stereo correspondence, winner-take-all, to disparity regularization is implemented fully on event-based hardware. Using a cluster of TrueNorth neurosynaptic processors, we demonstrate their ability to process bilateral event-based inputs streamed live by Dynamic Vision Sensors (DVS), at up to 2,000 disparity maps per second, producing high fidelity disparities which are in turn used to reconstru-ct, at low power, the depth of events produced from rapidly changing scenes. Experiments on real-world sequences demonstrate the ability of the system to take full advantage of the asyn-chronous and sparse nature of DVS sensors for low power depth reconstruction, in environments where conventional frame-based cameras connected to synchronous processors would be inefficient for rapidly moving objects. System evaluation on event-based sequences demonstrates a ~200X improvement in terms of power per pixel per disparity map compared to the closest state-of-the-art, and maximum latencies of up to 11ms from spike injection to disparity map ejection.
	1	Proceedings of the IEEE CVPR, 2019, pp. 989-997	Unsupervised Event-Based Learning of Optical Flow, Depth, and Egomotion	Alex Zihao Zhu, Liangzhe Yuan, Kenneth Chaney, Kostas Daniilidis	University of Pennsylvania	不相关	In this work, we propose a novel framework for unsupervised learning for event cameras that learns motion information from only the event stream. In particular, we propose an input representation of the events in the form of a discretized volume that maintains the temporal distribution of the events, which we pass through a neural network to predict the motion of the events. This motion is used to attempt to remove any motion blur in the event image. We then propose a loss function applied to the motion compensated event image that measures the motion blur in this image. We train two networks with this framework, one to predict optical flow, and one to predict egomotion and depths, and evaluate these networks on the Multi Vehicle Stereo Event Camera dataset, along with qualitative results from a variety of different scenes.
	2	Proceedings of the IEEE CVPR, 2019, pp. 1175-1186	Learning to Reconstruct People in Clothing From a Single RGB Camera	Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, Gerard Pons-Moll	Computer Graphics Lab, TU Braunschweig, Germany	相关(CNN)	We present Octopus, a learning-based model to infer the personalized 3D shape of people from a few frames (1-8) of a monocular video in which the person is moving with a reconstruction accuracy of 4 to 5mm, while being orders of magnitude faster than previous methods. From semantic segmentation images, our Octopus model reconstructs a 3D shape, including the parameters of SMPL plus clothing and hair in 10 seconds or less. The model achieves fast and accurate predictions based on two key design choices. First, by predicting shape in a canonical T-pose space, the network learns to encode the images of the person into pose-invariant latent codes, where the information is fused. Second, based on the observation that feed-forward predictions are fast but do not always align with the input images, we predict using both, bottom-up and top-down streams (one per view) allowing information to flow in both directions. Learning relies only on synthetic 3D data. Once learned, Octopus can take a variable number of frames as input, and is able to reconstruct shapes even from a single image with an accuracy of 5mm. Results on 3
	3	Proceedings of the IEEE CVPR, 2019, pp. 1197-1206	A Perceptual Prediction Framework for Self Supervised Event Segmentation	Sathyanarayanan N. Aakur, Sudeep Sarkar	University of South Florida Tampa, FL, USA	相关(CNN)	Temporal segmentation of long videos is an important problem, that has largely been tackled through supervised learning, often requiring large amounts of annotated training data. In this paper, we tackle the problem of self-supervised temporal segmentation that alleviates the need for any supervision in the form of labels (full supervision) or temporal ordering (weak supervi-sion). We introduce a self-supervised, predictive learning framework that draws inspiration from cognitive psychology to segment long, visually complex videos into constituent events. Learning involves only a single pass through the training data. We also introduce a new adaptive learning paradigm that helps reduce the effect of catastrophic forgetting in recurrent neural networks. Extensive experiments on three publicly available datasets - Breakfast Actions, 50 Salads, and INRIA Instructional Videos datasets show the efficacy of the proposed approach. We show that the proposed approach outperforms weakly-supervised and unsupervised baselines by up to 24% and achieves competitive segmentation results compared to fully supervised baselines with only a single pass through the training data. Finally, we show that the proposed self-supervised learning paradigm learns highly discriminating features to improve action recognition.
	4	Proceedings of the IEEE CVPR, 2019, pp. 1652-1660	Camera Lens Super-Resolution	Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, Feng Wu	University of Science and Technology of China	相关(CNN)	Existing methods for single image super-resolution (SR) are typically evaluated with synthetic degradation models such as bicubic or Gaussian downsampling. In this paper, we investigate SR from the perspective of camera lenses, named as CameraSR, which aims to alleviate the intrinsic tradeoff between resolution (R) and field-of-view (V) in realistic imaging systems. Specifically, we view the R-V degradation as a latent model in the SR process and learn to reverse it with realistic low- and high-resolution image pairs. To obtain the paired images, we propose two novel data acquisition strategies for two representative imaging systems (i.e., DSLR and smartphone cameras), respectively. Based on the obtained City100 dataset, we quantitatively analyze the performance of commonly-used synthetic degradation models, and demonstrate the superiority of CameraSR as a practical solution to boost the performance of existing SR methods. Moreover, CameraSR can be readily generalized to different content and devices, which serves as an advanced digital zoom tool in realistic imaging systems.
	5	Proceedings of the IEEE CVPR, 2019, pp. 3302-3312	Understanding the Limitations of CNN-Based Absolute Camera Pose Regression	Torsten Sattler, Qunjie Zhou, Marc Pollefeys, Laura Leal-Taixe	Chalmers University of Technology	相关(CNN)	Visual localization is the task of accurate camera pose estimation in a known scene. It is a key problem in computer vision and robotics, with applications including self-driving cars, Structure-from-Motion, SLAM, and Mixed Reality. Traditionally, the localization problem has been tackled using 3D geometry. Recently, end-to-end approaches based on convolutional neural networks have become popular. These methods learn to directly regress the camera pose from an input image. However, they do not achieve the same level of pose accuracy as 3D structure-based methods. To understand this behavior, we develop a theoretical model for camera pose regression. We use our model to predict failure cases for pose regression techniques and verify our predictions through experiments. We furthermore use our model to show that pose regression is more closely related to pose approximation via image retrieval than to accurate pose estimation via 3D structure. A key result is that current approaches do not consistently outperform a handcrafted image retrieval baseline. This clearly shows that additional research is needed before pose regression
	6	Proceedings of the IEEE CVPR, 2019, pp. 3743-3752	Polarimetric Camera Calibration Using an LCD Monitor	Zhixiang Wang, Yinqiang Zheng, Yung-Yu Chuang	National Taiwan University	不相关	It is crucial for polarimetric imaging to accurately calibrate the polarizer angles and the camera response function (CRF) of a polarizing camera. When this polarizing camera is used in a setting of multiview geometric imaging, it is often required to calibrate its intrinsic and extrinsic parameters as well, for which Zhang's calibration method is the most widely used with either a physical checker board, or more conveniently a virtual checker pattern displayed on a monitor. In this paper, we propose to jointly calibrate the polarizer angles and the inverse CRF (ICRF) using a slightly adapted checker pattern displayed on a liquid crystal display (LCD) monitor. Thanks to the lighting principles and the industry standards of the LCD monitors, the polarimetric and radiometric calibration can be significantly simplified, when assisted by the extrinsic parameters estimated from the checker pattern. We present a simple linear method for polarizer angle calibration and a convex method for radiometric calibration, both of which can be jointly refined in a process similar to bundle adjustment. Experiments have verified the feasibility and accuracy

7	Proceedings of the IEEE CVPR, 2019, pp. 3857-3866	Events-To-Video: Bringing Modern Computer Vision to Event Cameras	Henri Rebecq, Rene Ranftl, Vladlen Koltun, Davide Scaramuzza	/	不相关	Event cameras are novel sensors that report brightness changes in the form of asynchronous "events" instead of intensity frames. They have significant advantages over conventional cameras: high temporal resolution, high dynamic range, and no motion blur. Since the output of event cameras is fundamentally different from conventional cameras, it is commonly accepted that they require the development of specialized algorithms to accommodate the particular nature of events. In this work, we take a different view and propose to apply existing, mature computer vision techniques to videos reconstructed from event data. We propose a novel, recurrent neural network to reconstruct videos from a stream of events and train it on a large amount of simulated event data. Our experiments show that our approach surpasses state-of-the-art reconstruction methods by a large margin (> 20%) in terms of image quality. We further apply off-the-shelf computer vision algorithms to videos reconstructed from event data on tasks such as object classification and visual-inertial odometry, and show that this strategy consistently outperforms algorithms that were specifically designed for event data. We believe that our approach opens the door to bringing the outstanding properties of event cameras to an entirely new range of tasks.
8	Proceedings of the IEEE CVPR, 2019, pp. 3887-3896	EventNet: Asynchronous Recursive Event Processing	Yusuke Sekikawa, Kosuke Hara, Hideo Saito	Denso IT Laboratory	相关(CNN)	Event cameras are bio-inspired vision sensors that mimic retinas to asynchronously report per-pixel intensity changes rather than outputting an actual intensity image at regular intervals. This new paradigm of image sensor offers significant potential advantages; namely, sparse and non-redundant data representation. Unfortunately, however, most of the existing artificial neural network architectures, such as a CNN, require dense synchronous input data, and therefore, cannot make use of the sparseness of the data. We propose EventNet, a neural network designed for real-time processing of asynchronous event streams in a recursive and event-wise manner. EventNet models dependence of the output on tens of thousands of causal events recursively using a novel temporal coding scheme. As a result, at inference time, our network operates in an event-wise manner that is realized with very few sum-of-the-product operations---look-up table and temporal feature aggregation---which enables processing of 1 mega or more events per second on standard CPU. In experiments using real data, we demonstrated the real-time performance and robustness
9	Proceedings of the IEEE CVPR, 2019, pp. 5987-5995	Turn a Silicon Camera Into an InGaAs Camera	Feifan Lv, Yinqiang Zheng, Bohan Zhang, Feng Lu	State Key Laboratory of VR Technology and Systems, School of CSE, Beihang University, Beijing, China	相关(CNN)	Short-wave infrared (SWIR) imaging has a wide range of applications for both industry and civilian. However, the InGaAs sensors commonly used for SWIR imaging suffer from a variety of drawbacks, including high price, low resolution, unstable quality, and so on. In this paper, we propose a novel solution for SWIR imaging using a common Silicon sensor, which has cheaper price, higher resolution and better technical maturity compared with the specialized InGaAs sensor. Our key idea is to approximate the response of the InGaAs sensor by exploiting the largely ignored sensitivity of a Silicon sensor, weak as it is, in the SWIR range. To this end, we build a multi-channel optical system to collect a new SWIR dataset and present a physically meaningful three-stage image processing algorithm on the basis of CNN. Both qualitative and quantitative experiments show promising experimental results, which demonstrate the effectiveness of the proposed method.
10	Proceedings of the IEEE CVPR, 2019, pp. 6358-6367	EV-Gait: Event-Based Robust Gait Recognition Using Dynamic Vision Sensors	Yanxiang Wang, Bowen Du, Yiran Shen, Kai Wu, Guangrong Zhao, Jianguo Sun, Hongkai Wen	Harbin Engineering University	相关(DNN)	In this paper, we introduce a new type of sensing modality, the Dynamic Vision Sensors (Event Cameras), for the task of gait recognition. Compared with the traditional RGB sensors, the event cameras have many unique advantages such as ultra low resources consumption, high temporal resolution and much larger dynamic range. However, those cameras only produce noisy and asynchronous events of intensity changes rather than frames, where conventional vision-based gait recognition algorithms can't be directly applied. To address this, we propose a new Event-based Gait Recognition (EV-Gait) approach, which exploits motion consistency to effectively remove noise, and uses a deep neural network to recognise gait from the event streams. To evaluate the performance of EV-Gait, we collect two event-based gait datasets, one from real-world experiments and the other by converting the publicly available RGB gait recognition benchmark CASIA-B. Extensive experiments show that EV-Gait can get nearly 96% recognition accuracy in the real-world settings, while on the CASIA-B benchmark it achieves comparable performance with state-of-
11	Proceedings of the IEEE CVPR, 2019, pp. 6820-6829	Bringing a Blurry Frame Alive at High Frame-Rate With an Event Camera	Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, Yuchao Dai	Australian National University, Canberra, Australia	相关(CNN)	Event-based cameras can measure intensity changes (called 'events') with microsecond accuracy under high-speed motion and challenging lighting conditions. With the active pixel sensor (APS), the event camera allows simultaneous output of the intensity frames. However, the output images are captured at a relatively low frame-rate and often suffer from motion blur. A blurry image can be regarded as the integral of a sequence of latent images, while the events indicate the changes between the latent images. Therefore, we are able to model the blur-generation process by associating event data to a latent image. In this paper, we propose a simple and effective approach, the Event-based Double Integral (EDI) model, to reconstruct a high frame-rate, sharp video from a single blurry frame and its event data. The video generation is based on solving a simple non-convex optimization problem in a single scalar variable. Experimental results on both synthetic and real images demonstrate the superiority of our EDI model and optimization method in comparison to the state-of-the-art.
12	Proceedings of the IEEE CVPR, 2019, pp. 7842-7851	Object-Centric Auto-Encoders and Dummy Anomalies for Abnormal Event Detection in Video	Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, Ling Shao	Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, UAE	相关(CNN)	Abnormal event detection in video is a challenging vision problem. Most existing approaches formulate abnormal event detection as an outlier detection task, due to the scarcity of anomalous data during training. Because of the lack of prior information regarding abnormal events, these methods are not fully-equipped to differentiate between normal and abnormal events. In this work, we formalize abnormal event detection as a one-versus-rest binary classification problem. Our contribution is two-fold. First, we introduce an unsupervised feature learning framework based on object-centric convolutional auto-encoders to encode both motion and appearance information. Second, we propose a supervised classification approach based on clustering the training samples into normality clusters. A one-versus-rest abnormal event classifier is then employed to separate each normality cluster from the rest. For the purpose of training the classifier, the other clusters act as dummy anomalies. During inference, an object is labeled as abnormal if the highest classification score assigned by the one-versus-rest classifiers is negative. Comprehensive experiments are performed on four benchmarks: Avenue, ShanghaiTech, UCSD and UMN. Our approach provides superior results on all four data sets. On the large-scale ShanghaiTech data set, our method provides an absolute gain of 8.4% in terms of frame-level AUC compared to the state-of-the-art method.
13	Proceedings of the IEEE CVPR, 2019, pp. 8730-8738	What Correspondences Reveal About Unknown Camera and Motion Models?	Thomas Probst, Ajad Chhatkuli, Danda Pani Paudel, Luc Van Gool	Computer Vision Laboratory, ETH Zurich, Switzerland	不相关	In two-view geometry, camera models and motion types are used as key knowledge along with the image point correspondences in order to solve several key problems of 3D vision. Problems such as Structure-from-Motion (SfM) and camera self-calibration are tackled under the assumptions of a specific camera projection model and motion type. However, these key assumptions may not be always justified, i.e.., we may often know neither the camera model nor the motion type beforehand. In that context, one can extract only the point correspondences between images. From such correspondences, recovering two-view relationship --expressed by the unknown camera model and motion type-- remains to be an unsolved problem. In this paper, we tackle this problem in two steps. First, we propose a method that computes the correct two-view relationship in the presence of noise and outliers. Later, we study different possibilities to disambiguate the obtained relationships into camera model and motion type. By extensive experiments on both synthetic and real data, we verify our theory and assumptions in practical settings.

2019	14	Proceedings of the IEEE CVPR, 2019, pp. 8797-8806	CityFlow: A City-Scale Benchmark for Multi-Target Multi-Camera Vehicle Tracking and Re-Identification	Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, Jenq-Neng Hwang	University of Washington	不相关	Urban traffic optimization using traffic cameras as sensors is driving the need to advance state-of-the-art multi-target multi-camera (MTMC) tracking. This work introduces CityFlow, a city-scale traffic camera dataset consisting of more than 3 hours of synchronized HD videos from 40 cameras across 10 intersections, with the longest distance between two simultaneous cameras being 2.5 km. To the best of our knowledge, CityFlow is the largest-scale dataset in terms of spatial coverage and the number of cameras/videos in an urban environment. The dataset contains more than 200K annotated bounding boxes covering a wide range of scenes, viewing angles, vehicle models, and urban traffic flow conditions. Camera geometry and calibration information are provided to aid spatio-temporal analysis. In addition, a subset of the benchmark is made available for the task of image-based vehicle re-identification (ReID). We conducted an extensive experimental evaluation of baselines/state-of-the-art approaches in MTMC tracking, multi-target single-camera (MTSC) tracking, object detection, and image-based ReID on this dataset, analyzing the impact of different network architectures, loss functions, spatio-temporal models and their combinations on task effectiveness. An evaluation server is launched with the release of our benchmark at the 2019 AI City Challenge (https://www.aicitychallenge.org/) that allows researchers to compare the performance of their newest techniques. We expect this dataset to catalyze research in this field, propel the state-of-the-art forward, and lead to deployed traffic optimization(s) in the real world.
	15	Proceedings of the IEEE CVPR, 2019, pp. 9327-9335	Rare Event Detection Using Disentangled Representation Learning	Ryuhei Hamaguchi, Ken Sakurada, Ryosuke Nakamura	National Institute of Advanced Industrial Science and Technology (AIST)	不相关	This paper presents a novel method for rare event detection from an image pair with class-imbalanced datasets. A straightforward approach for event detection tasks is to train a detection network from a large-scale dataset in an end-to-end manner. However, in many applications such as building change detection on satellite images, few positive samples are available for the training. Moreover, an image pair of scenes contains many trivial events, such as in illumination changes or background motions. These many trivial events and the class imbalance problem lead to false alarms for rare event detection. In order to overcome these difficulties, we propose a novel method to learn disentangled representations from only low-cost negative samples. The proposed method disentangles the different aspects in a pair of observations: variant and invariant factors that represent trivial events and image contents, respectively. The effective-ness of the proposed approach is verified by the quantitative evaluations on four change detection datasets, and the qualitative analysis shows that the proposed method can acquire the representations that disentangle
	16	Proceedings of the IEEE CVPR, 2019, pp. 9709-9718	Volumetric Capture of Humans With a Single RGBD Camera via Semi-Parametric Learning	Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, Shahram Izadi, Sean Fanello	Google Inc	不相关	Volumetric (4D) performance capture is fundamental for AR/VR content generation. Whereas previous work in 4D performance capture has shown impressive results in studio settings, the technology is still far from being accessible to a typical consumer who, at best, might own a single RGBD sensor. Thus, in this work, we propose a method to synthesize free viewpoint renderings using a single RGBD camera. The key insight is to leverage previously seen "calibration" images of a given user to extrapolate what should be rendered in a novel viewpoint from the data available in the sensor. Given these past observations from multiple viewpoints, and the current RGBD image from a fixed view, we propose an end-to-end framework that fuses both these data sources to generate novel renderings of the performer. We demonstrate that the method can produce high fidelity images, and handle extreme changes in subject pose and camera viewpoints. We also show that the system generalizes to performers not seen in the training data. We run exhaustive experiments demonstrating the effectiveness of the proposed semi-parametric model (i.e. calibration images available to the neural network) compared to other state of the art machine learned solutions. Further, we compare the method with more traditional pipelines that employ multi-view capture. We show that our framework is able to achieve compelling results, with substantially less infrastructure than previously required.
	17	Proceedings of the IEEE CVPR, 2019, pp. 10081-10090	Event-Based High Dynamic Range Image and Very High Frame Rate Video Generation Using Conditional Generative Adversarial Networks	Lin Wang, S. Mohammad Mostafavi I., Yo-Sung Ho, Kuk-Jin Yoon	Visual Intelligence Laboratory, Dept. Mechanical Engineering, KAIST, Korea	相关(CNN)	Event cameras have a lot of advantages over traditional cameras, such as low latency, high temporal resolution, and high dynamic range. However, since the outputs of event cameras are the sequences of asynchronous events over time rather than actual intensity images, existing algorithms could not be directly applied. Therefore, it is demanding to generate intensity images from events for other tasks. In this paper, we unlock the potential of event camera-based conditional generative adversarial networks to create images/videos from an adjustable portion of the event data stream. The stacks of space-time coordinates of events are used as inputs and the network is trained to reproduce images based on the spatio-temporal intensity changes. The usefulness of event cameras to generate high dynamic range (HDR) images even in extreme illumination conditions and also non blurred images under rapid motion is also shown. In addition, the possibility of generating very high frame rate videos is demonstrated, theoretically up to 1 million frames per second(FPS) since the temporal resolution of event cameras is about 1 microsecond. Proposed methods are evaluated by comparing the results with the intensity images captured on the same pixel grid-line of events using online available real datasets and synthetic datasets produced by the
	18	Proceedings of the IEEE CVPR, 2019, pp. 10121-10129	Ray-Space Projection Model for Light Field Camera	Qi Zhang, Jinbo Ling, Qing Wang, Jingyi Yu	School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, P.R. China,	相关	Light field essentially represents the collection of rays in space. The rays captured by multiple light field cameras form subsets of full rays in 3D space and can be transformed to each other. However, most previous approaches model the projection from an arbitrary point in 3D space to corresponding pixel on the sensor. There are few models on describing the ray sampling and transformation among multiple light field cameras. In the paper, we propose a novel ray-space projection model to transform sets of rays captured by multiple light field cameras in term of the Plucker coordinates. We first derive a 6x6 ray-space intrinsic matrix based on multi-projection-center (MPC) model. A homogeneous ray-space projection matrix and a fundamental matrix are then proposed to establish ray-ray correspondences among multiple light fields. Finally, based on the ray-space projection matrix, a novel camera calibration method is proposed to verify the proposed model. A linear constraint and a ray-ray cost function are established for linear initial solution and non-linear optimization respectively. Experimental results on both synthetic and real light field data have verified the effectiveness and robustness of the proposed model.
	19	Proceedings of the IEEE CVPR, 2019, pp. 10245-10254	Speed Invariant Time Surface for Learning to Detect Corner Points With Event-Based Cameras	Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, Vincent Lepetit	Prophesee, Paris, France	不相关	We propose a learning approach to corner detection for event-based cameras that is stable even under fast and abrupt motions. Event-based cameras offer high temporal resolution, power efficiency, and high dynamic range. However, the properties of event-based data are very different compared to standard intensity images, and simple extensions of corner detection methods designed for these images do not perform well on event-based data. We first introduce an efficient way to compute a time surface that is invariant to the speed of the objects. We then show that we can train a Random Forest to recognize events generated by a moving corner from our time surface. Random Forests are also extremely efficient, and therefore a good choice to deal with the high capture frequency of event-based cameras ---our implementation processes up to 1.6Mev/s on a single CPU. Thanks to our time surface formulation and this learning approach, our method is significantly more robust to abrupt changes of direction of the corners compared to previous ones. Our method also naturally assigns a confidence score for the corners, which can be useful for postprocessing. Moreover, we introduce a high-resolution dataset suitable for quantitative evaluation and comparison of corner detection methods for event-based cameras. We call our approach SILC, for Speed Invariant Learned Corners, and compare it to the state-of-the-art with extensive experiments, showing better performance.
	20	Proceedings of the IEEE CVPR, 2019, pp. 10986-10995	Neural RGB(r)D Sensing: Depth and Uncertainty From a Video Camera	Chao Liu, Jinwei Gu, Kihwan Kim, Srinivasa G. Narasimhan, Jan Kautz	NVIDIA	相关	Depth sensing is crucial for 3D reconstruction and scene understanding. Active depth sensors provide dense metric measurements, but often suffer from limitations such as restricted operat-ing ranges, low spatial resolution, sensor interference, and high power consumption. In this paper, we propose a deep learning (DL) method to estimate per-pixel depth and its uncerta-inty continuously from a monocular video stream, with the goal of effectively turning an RGB camera into an RGB-D camera. Unlike prior DL-based methods, we estimate a depth probabili-ty distribution for each pixel rather than a single depth value, leading to an estimate of a 3D depth probability volume for each input frame. These depth probability volumes are accumulated over time under a Bayesian filtering framework as more incoming frames are processed sequentially, which effectively reduces depth uncertainty and improves accuracy, robustness, and temporal stability. Compared to prior work, the proposed approach achieves more accurate and stable results, and generalizes better to new datasets. Experimental results also show the output of our approach can be directly fed into classical RGB-D based 3D scanning methods for 3D scene reconstruction.

	21	Proceedings of the IEEE CVPR, 2019, pp. 11796-11806	The Alignment of the Spheres: Globally-Optimal Spherical Mixture Alignment for Camera Pose Estimation	Dylan Campbell, Lars Petersson, Laurent Kneip, Hongdong Li, Stephen Gould	Australian National University	不相关	Determining the position and orientation of a calibrated camera from a single image with respect to a 3D model is an essential task for many applications. When 2D-3D correspondences can be obtained reliably, perspective-n-point solvers can be used to recover the camera pose. However, without the pose it is non-trivial to find cross-modality correspondences between 2D images and 3D models, particularly when the latter only contains geometric information. Consequently, the problem becomes one of estimating pose and correspondences jointly. Since outliers and local optima are so prevalent, robust objective functions and global search strategies are desirable. Hence, we cast the problem as a 2D-3D mixture model alignment task and propose the first globally-optimal solution to this formulation under the robust L2 distance between mixture distributions. We derive novel bounds on this objective function and employ branch-and-bound to search the 6D space of camera poses, guaranteeing global optimality without requiring a pose estimate. To accelerate convergence, we integrate local optimization, implement GPU bound computations, and provide an intuitive way to incorporate side information such as semantic labels. The algorithm is evaluated on challenging synthetic and real datasets, outperforming existing approaches and reliably converging to the global optimum.
	22	Proceedings of the IEEE CVPR, 2019, pp. 11817-11825	Deep Single Image Camera Calibration With Radial Distortion	Manuel Lopez, Roger Mari, Pau Gargallo, Yubin Kuang, Javier Gonzalez-Jimenez, Gloria Haro	/	不相关	Single image calibration is the problem of predicting the camera parameters from one image. This problem is of importance when dealing with images collected in uncontrolled conditions by non-calibrated cameras, such as crowd-sourced applications. In this work we propose a method to predict extrinsic (tilt and roll) and intrinsic (focal length and radial distortion) parameters from a single image. We propose a parameterization for radial distortion that is better suited for learning than directly predicting the distortion parameters. Moreover, predicting additional heterogeneous variables exacerbates the problem of loss balancing. We propose a new loss function based on point projections to avoid having to balance heterogeneous loss terms. Our method is, to our knowledge, the first to jointly estimate the tilt, roll, focal length, and radial distortion parameters from a single image. We thoroughly analyze the performance of the proposed method and the impact of the improvements and compare with previous approaches for single image radial distortion correction.
	23	Proceedings of the IEEE CVPR, 2019, pp. 11826-11835	CAM-Convs: Camera-Aware Multi-Scale Convolutions for Single-View Depth	Jose M. Facil, Benjamin Ummenhofer, Huizhong Zhou, Luis Montesano, Thomas Brox, Javier Civera	University of Zaragoza	相关(CNN)	Single-view depth estimation suffers from the problem that a network trained on images from one camera does not generalize to images taken with a different camera model. Thus, changing the camera model requires collecting an entirely new training dataset. In this work, we propose a new type of convolution that can take the camera parameters into account, thus allowing neural networks to learn calibration-aware patterns. Experiments confirm that this improves the generalization capabilities of depth prediction networks considerably, and clearly outperforms the state of the art when the train and test images are acquired with different cameras.
	24	Proceedings of the IEEE CVPR, 2019, pp. 12240-12249	Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation	Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, Michael J. Black	Max Planck Institute for Intelligent Systems	相关	We address the unsupervised learning of several interconnected problems in low-level vision: single view depth prediction, camera motion estimation, optical flow, and segmentation of a video into the static scene and moving regions. Our key insight is that these four fundamental vision problems are coupled through geometric constraints. Consequently, learning to solve them together simplifies the problem because the solutions can reinforce each other. We go beyond previous work by exploiting geometry more explicitly and segmenting the scene into static and moving regions. To that end, we introduce Competitive Collaboration, a framework that facilitates the coordinated training of multiple specialized neural networks to solve complex problems. Competitive Collaboration works much like expectation-maximization, but with neural networks that act as both competitors to explain pixels that correspond to static or moving regions, and as collaborators through a moderator that assigns pixels to be either static or independently moving. Our novel method integrates all these problems in a common framework and simultan-eously reasons about the segmentation of the scene into moving objects and the static background, the camera motion, depth of the static scene structure, and the optical flow of moving objects. Our model is trained without any supervision and achieves state-of-the-art performance among joint unsupervised methods on all sub-problems.
	25	Proceedings of the IEEE CVPR, 2019, pp. 12280-12289	Focus Is All You Need: Loss Functions for Event-Based Vision	Guillermo Gallego, Mathias Gehrig, Davide Scaramuzza	/	不相关	Event cameras are novel vision sensors that output pixel-level brightness changes ("events") instead of traditional video frames. These asynchronous sensors offer several advantages over traditional cameras, such as, high temporal resolution, very high dynamic range, and no motion blur. To unlock the potential of such sensors, motion compensation methods have been recently proposed. We present a collection and taxonomy of twenty two objective functions to analyze event alignment in motion compensation approaches. We call them focus loss functions since they have strong connections with functions used in traditional shape-from-focus applications. The proposed loss functions allow bringing mature computer vision tools to the realm of event cameras. We compare the accuracy and runtime performance of all loss functions on a publicly available dataset, and conclude that the variance, the gradient and the Laplacian magnitudes are among the best loss functions. The applicability of the loss functions is shown on multiple tasks: rotational motion, depth and optical flow estimation. The proposed focus loss functions allow to unlock the
	26	Proceedings of the IEEE CVPR, 2019, pp. 12300-12308	Event Cameras, Contrast Maximization and Reward Functions: An Analysis	Timo Stoffregen, Lindsay Kleeman	Dept. Electrical and Computer Systems Engineering, Monash University, Australia	不相关	Event cameras asynchronously report timestamped changes in pixel intensity and offer advantages over conventional raster scan cameras in terms of low-latency, low redundancy sensing and high dynamic range. In recent years, much of research in event based vision has been focused on performing tasks such as optic flow estimation, moving object segmentation, feature tracking, camera rotation estimation and more, through contrast maximization. In contrast maximization, events are warped along motion trajectories whose parameters depend on the quantity being estimated, to some time t_ref. The parameters are then scored by some reward function of the accumulated events at t_ref. The versatility of this approach has lead to a flurry of research in recent years, but no in-depth study of the reward chosen during optimization has yet been made. In this work we examine the choice of reward used in contrast maxim-ization, propose a classification of different rewards and show how a reward can be constructed that is more robust to noise and aperture uncertainty. We validate our work experimentally by predicting optical flow and comparing
	1	Proceedings of the IEEE CVPR, 2020, pp. 6040-6049	Lightweight Multi-View 3D Pose Estimation Through Camera-Disentangled Representation	Edoardo Remelli, Shangchen Han, Sina Honari, Pascal Fua, Robert Wang	CVLab, EPFL, Lausanne, Switzerland	相关(CNN)	We present a lightweight solution to recover 3D pose from multi-view images captured with spatially calibrated cameras. Building upon recent advances in interpretable representation learning, we exploit 3D geometry to fuse input images into a unified latent representation of pose, which is disentangled from camera view-points. This allows us to reason effectively about 3D pose across different views without using compute-intensive volumetric grids. Our architecture then conditions the learned representation on camera projection operators to produce accurate per-view 2d detections, that can be simply lifted to 3D via a differentiable Direct Linear Transform (DLT) layer. In order to do it efficiently, we propose a novel implementation of DLT that is orders of magnitude faster on GPU architectures than standard SVD-based triangulation methods. We evaluate our approach on two large-scale human pose datasets (H36M and Total Capture): our method outperforms or performs comparably to the state-of-the-art volumetric methods, while, unlike them, yielding real-time performance.
	2	Proceedings of the IEEE CVPR, 2020, pp. 3586-3595	Video to Events: Recycling Video Datasets for Event Cameras	Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrio, Davide Scaramuzza	Dept. Informatics, Univ. of Zurich and Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich	相关(CNN)	Event cameras are novel sensors that output brightness changes in the form of a stream of asynchronous "events" instead of intensity frames. They offer significant advantages with respect to conventional cameras: high dynamic range (HDR), high temporal resolution, and no motion blur. Recently, novel learning approaches operating on event data have achieved impressive results. Yet, these methods require a large amount of event data for training, which is hardly available due the novelty of event sensors in computer vision research. In this paper, we present a method that addresses these needs by converting any existing video dataset recorded with conventional cameras to synthetic event data. This unlocks the use of a virtually unlimited number of existing video datasets for training networks designed for real event data. We evaluate our method on two relevant vision tasks, i.e., object recognition and semantic segmentation, and show that models trained on synthetic events have several benefits: (i) they generalize well to real event data, even in scenarios where standard-camera images are blurry or overexpos-ed, by inheriting the outstanding properties of event cameras; (ii) they can be used for fine-tuning on real data to improve over state-of-the-art for both classification and semantic segmentation.

3	Proceedings of the IEEE CVPR, 2020, pp. 2535-2544	Why Having 10,000 Parameters in Your Camera Model Is Better Than Twelve	Thomas Schops, Viktor Larsson, Marc Pollefeys, Torsten Sattler	Department of Computer Science, ETH Zurich	不相关	Camera calibration is an essential first step in setting up 3D Computer Vision systems. Commonly used parametric camera models are limited to a few degrees of freedom and thus often do not optimally fit to complex real lens distortion. In contrast, generic camera models allow for very accurate calibration due to their flexibility. Despite this, they have seen little use in practice. In this paper, we argue that this should change. We propose a calibration pipeline for generic models that is fully automated, easy to use, and can act as a drop-in replacement for parametric calibration, with a focus on accuracy. We compare our results to parametric calibrations. Considering stereo depth estimation and camera pose estimation as examples, we show that the calibration error acts as a bias on the results. We thus argue that in contrast to current common practice, generic models should be preferred over parametric ones whenever possible. To facilitate this, we released our calibration pipeline at https://github.com/puzzlepaint/camera_calibration, making both easy-to-use and accurate camera calibration available to everyone.
4	Proceedings of the IEEE CVPR, 2020, pp. 2950-2959	Camera Trace Erasing	Chang Chen, Zhiwei Xiong, Xiaoming Liu, Feng Wu	University of Science and Technology of China	相关(CNN)	Camera trace is a unique noise produced in digital imaging process. Most existing forensic methods analyze camera trace to identify image origins. In this paper, we address a new low-level vision problem, camera trace erasing, to reveal the weakness of trace-based forensic methods. A comprehensive investigation on existing anti-forensic methods reveals that it is non-trivial to effectively erase camera trace while avoiding the destruction of content signal. To reconcile these two demands, we propose Siamese Trace Erasing (SiamTE), in which a novel hybrid loss is designed on the basis of Siamese architecture for network training. Specifically, we propose embedded similarity, truncated fidelity, and cross identity to form the hybrid loss. Compared with existing anti-forensic methods, SiamTE has a clear advantage for camera trace erasing, which is demonstrated in three representative tasks.
5	Proceedings of the IEEE CVPR, 2020, pp. 3022-3032	What Does Plate Glass Reveal About Camera Calibration?	Qian Zheng, Jinnan Chen, Zhan Lu, Boxin Shi, Xudong Jiang, Kim-Hui Yap, Ling-Yu Duan, Alex C. Kot	School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore	不相关	This paper aims to calibrate the orientation of glass and the field of view of the camera from a single reflection-contaminated image. We show how a reflective amplitude coefficient map can be used as a calibration cue. Different from existing methods, the proposed solution is free from image contents. To reduce the impact of a noisy calibration cue estimated from a reflection-contaminated image, we propose two strategies: an optimization-based method that imposes part of though reliable entries on the map and a learning-based method that fully exploits all entries. We collect a dataset containing 320 samples as well as their camera parameters for evaluation. We demonstrate that our method not only facilitates a general single image camera calibration method that leverages image contents but also contributes to improving the performance of single image reflection removal. Furthermore, we show our byproduct output helps alleviate the ill-posed problem of estimating the panorama from a single image.
6	Proceedings of the IEEE CVPR, 2020, pp. 7173-7182	Can Facial Pose and Expression Be Separated With Weak Perspective Camera?	Evangelos Sariyanidi, Casey J. Zampella, Robert T. Schultz, Birkan Tunc	Center for Autism Research, Children's Hospital of Philadelphia	不相关	Separating facial pose and expression within images requires a camera model for 3D-to-2D mapping. The weak perspective (WP) camera has been the most popular choice; it is the default, if not the only option, in state-of-the-art facial analysis methods and software. WP camera is justified by the supposition that its errors are negligible when the subjects are relatively far from the camera, yet this claim has never been tested despite nearly 20 years of research. This paper critically examines the suitability of WP camera for separating facial pose and expression. First, we theoretically show that WP causes pose-expression ambiguity, as it leads to estimation of spurious expressions. Next, we experimentally quantify the magnitude of spurious express-ions. Finally, we test whether spurious expressions have detrimental effects on a common facial analysis application, namely Action Unit (AU) detection. Contrary to conventional wisdom, we find that severe pose-expression ambiguity exists even when subjects are not close to the camera, leading to large false positive rates in AU detection. We also demonstrate that the magnitude and characteristics of spurious expressions depend on the point distribution model used to model the expressions. Our results suggest that common assumptions about WP need to be revisited in facial expression modeling, and that facial analysis software should encourage and facilitate the use of the true camera model whenever possible.
7	Proceedings of the IEEE CVPR, 2020, pp. 5336-5345	Novel View Synthesis of Dynamic Scenes With Globally Coherent Depths From a Monocular Camera	Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, Jan Kautz	University of Minnesota	不相关	This paper presents a new method to synthesize an image from arbitrary views and times given a collection of images of a dynamic scene. A key challenge for the novel view synthesis arises from dynamic scene reconstruction where epipolar geometry does not apply to the local motion of dynamic contents. To address this challenge, we propose to combine the depth from single view (DSV) and the depth from multi-view stereo (DMV), where DSV is complete, i.e., a depth is assigned to every pixel, yet view-variant in its scale, while DMV is view-invariant yet incomplete. Our insight is that although its scale and quality are inconsistent with other views, the depth estimation from a single view can be used to reason about the globally coherent geometry of dynamic contents. We cast this problem as learning to correct the scale of DSV, and to refine each depth with locally consistent motions between views to form a coherent depth estimation. We integrate these tasks into a depth fusion network in a self-supervised fashion. Given the fused depth maps, we synthesize a photorealistic virtual view in a specific location and time with our deep blending network that completes the scene and renders the virtual view. We evaluate our method of depth estimation and view synthesis on a diverse real-world dynamic scenes and show the outstanding performance over existing methods.
8	Proceedings of the IEEE CVPR, 2020, pp. 14414-14423	Learning Visual Motion Segmentation Using Event Surfaces	Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, Yiannis Aloimonos	University of Maryland, College Park College Park, Maryland	不相关	Event-based cameras have been designed for scene motion perception - their high temporal resolution and spatial data sparsity converts the scene into a volume of boundary trajectories and allows to track and analyze the evolution of the scene in time. Analyzing this data is computationally expensive, and there is substantial lack of theory on dense-in-time object motion to guide the development of new algorithms; hence, many works resort to a simple solution of discretizing the event stream and converting it to classical pixel maps, which allows for application of conventional image processing methods. In this work we present a Graph Convolutional neural network for the task of scene motion segmentation by a moving camera. We convert the event stream into a 3D graph in (x,y,t) space and keep per-event temporal information. The difficulty of the task stems from the fact that unlike in metric space, the shape of an object in (x,y,t) space depends on its motion and is not the same across the dataset. We discuss properties of of the event data with respect to this 3D recognition problem, and show that our Graph Convolutional architecture is superior to PointNet++. We evaluate our method on the state of the art event-based motion segmentation dataset - EV-IMO and perform comparisons to a frame-based method proposed by its authors. Our ablation studies show that increasing the event slice width improves the accuracy, and how subsampling and edge configurations affect the network performance.
9	Proceedings of the IEEE CVPR, 2020, pp. 1730-1739	Neuromorphic Camera Guided High Dynamic Range Imaging	Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, Boxin Shi	Key Laboratory of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University	相关(CNN)	Reconstruction of high dynamic range image from a single low dynamic range image captured by a frame-based conventional camera, which suffers from over- or under-exposure, is an ill-posed problem. In contrast, recent neuromorphic cameras are able to record high dynamic range scenes in the form of an intensity map, with much lower spatial resolution, and without color. In this paper, we propose a neuromorphic camera guided high dynamic range imaging pipeline, and a network consisting of specially designed modules according to each step in the pipeline, which bridges the domain gaps on resolution, dynamic range, and color representation between two types of sensors and images. A hybrid camera system has been built to validate that the proposed method is able to reconstruct quantitatively and qualitatively high-quality high dynamic range images by successfully fusing the images and intensity maps for various real-world scenarios.
10	Proceedings of the IEEE CVPR, 2020, pp. 12585-12594	JA-POLS: A Moving-Camera Background Model via Joint Alignment and Partially-Overlapping Local Subspaces	Irit Chelly, Vlad Winter, Dor Litvak, David Rosen, Oren Freifeld	Ben-Gurion University	不相关	Background models are widely used in computer vision. While successful Static-camera Background (SCB) models exist, Moving-camera Background (MCB) models are limited. Seeming-ly, there is a straightforward solution: 1) align the video frames; 2) learn an SCB model; 3) warp either original or previously-unseen frames toward the model. This approach, however, has drawbacks, especially when the accumulative camera motion is large and/or the video is long. Here we propose a purely-2D unsupervised modular method that systematically eliminates those issues. First, to estimate warps in the original video, we solve a joint-alignment problem while leveraging a certifiably-correct initialization. Next, we learn both multiple partially-overlapping local subspaces and how to predict alignments. Lastly, in test time, we warp a previously-unseen frame, based on the prediction, and project it on a subset of those subspaces to obtain a background/foreground separation. We show the method handles even large scenes with a relatively-free camera motion (provided the camera-to-scene distance does not change much) and that it not only yields State-of-the-Art results on the original video but also generalizes gracefully to previously-unseen videos of the same scene. Our code is available at https://github.com/BGU-CS-VIL/JA-POLS.

2020	11	Proceedings of the IEEE CVPR, 2020, pp. 2545-2554	Blur Aware Calibration of Multi-Focus Plenoptic Camera	Mathieu Labussiere, Celine Teuliere, Frederic Bernardin, Omar Ait-Aider	Université Clermont Auvergne, CNRS, SIGMA Clermont	不相关	This paper presents a novel calibration algorithm for Multi-Focus Plenoptic Cameras (MFPCs) using raw images only. The design of such cameras is usually complex and relies on precise placement of optic elements. Several calibration procedures have been proposed to retrieve the camera parameters but relying on simplified models, reconstructed images to extract features, or multiple calibrations when several types of micro-lens are used. Considering blur information, we propose a new Blur Aware Plenoptic (BAP) feature. It is first exploited in a precalibra-tion step that retrieves initial camera parameters, and secondly to express a new cost function for our single optimization process. The effectiveness of our calibration method is validated by quantitative and qualitative experiments.
	12	Proceedings of the IEEE CVPR, 2020, pp. 1651-1660	Single-Image HDR Reconstruction by Learning to Reverse the Camera Pipeline	Yulun Liu, Weisheng Lai, Yusheng Chen, Yilung Kao, Minghsuan Yang, Yungyu Chuang, Jiabin Huang	National Taiwan University	相关(CNN)	Recovering a high dynamic range (HDR) image from a single low dynamic range (LDR) input image is challenging due to missing details in under-/over-exposed regions caused by quantization and saturation of camera sensors. In contrast to existing learning-based methods, our core idea is to incorporate the domain knowledge of the LDR image formation pipeline into our model. We model the HDR-to-LDR image formation pipeline as the (1) dynamic range clipping, (2) non-linear mapping from a camera response function, and (3) quantization. We then propose to learn three specialized CNNs to reverse these steps. By decomposing the problem into specific sub-tasks, we impose effective physical constraints to facilitate the training of individual sub-networks. Finally, we jointly fine-tune the entire model end-to-end to reduce error accumulation. With extensive quantitative and qualitative experiments on diverse image datasets, we demonstrate that the proposed method performs favorably against state-of-the-art single-image HDR reconstruction algorithms.
	13	Proceedings of the IEEE CVPR, 2020, pp. 6021-6030	Averaging Essential and Fundamental Matrices in Collinear Camera Settings	Amnon Geifman, Yoni Kasten, Meirav Galun, Ronen Basri	Weizmann Institute of Science	不相关	Global methods to Structure from Motion have gained popularity in recent years. A significant drawback of global methods is their sensitivity to collinear camera settings. In this paper, we introduce an analysis and algorithms for averaging bifocal tensors (essential or fundamental matrices) when either subsets or all of the camera centers are collinear. We provide a complete spectral characterization of bifocal tensors in collinear scenarios and further propose two averaging algorithms. The first algorithm uses rank constrained minimization to recover camera matrices in fully collinear settings. The second algorithm enriches the set of possibly mixed collinear and non-collinear cameras with additional, "virtual cameras," which are placed in general position, enabling the application of existing averaging methods to the enriched set of bifocal tensors. Our algorithms are shown to achieve state of the art results on various benchmarks that include autonomous car datasets and unordered image collections in both calibrated and unclibrated settings.
	14	Proceedings of the IEEE CVPR, 2020, pp. 7529-7538	Hardware-in-the-Loop End-to-End Optimization of Camera Image Processing Pipelines	Ali Mosleh, Avinash Sharma, Emmanuel Onzon, Fahim Mannan, Nicolas Robidoux, Felix Heide	Algolux	不相关	Commodity imaging systems rely on hardware image signal processing (ISP) pipelines. These low-level pipelines consist of a sequence of processing blocks that, depending on their hyperp-arameters, reconstruct a color image from RAW sensor measurements. Hardware ISP hyperparameters have a complex interaction with the output image, and therefore with the down-stream application ingesting these images. Traditionally, ISPs are manually tuned in isolation by imaging experts without an end-to-end objective. Very recently, ISPs have been optimized with 1st-order methods that require differentiable approximations of the hardware ISP. Departing from such approximations, we present a hardware-in-the-loop method that directly optimi-zes hardware image processing pipelines for end-to-end domain-specific losses by solving a nonlinear multi-objective optimization problem with a novel 0th-order stochastic solver directly interfaced with the hardware ISP. We validate the proposed method with recent hardware ISPs and 2D object detection, segmentation, and human viewing as end-to-end downstream tasks. For automotive 2D object detection, the proposed method outperforms manual expert tuning by 30% mean average precision (mAP) and recent methods using ISP approximations by 18% mAP.
	15	Proceedings of the IEEE CVPR, 2020, pp. 1672-1681	Single Image Optical Flow Estimation With an Event Camera	Liyuan Pan, Miaomiao Liu, Richard Hartley	Australian National University, Canberra, Australia	相关(CNN)	Event cameras are bio-inspired sensors that asynchronously report intensity changes in microsecond resolution. DAVIS can capture high dynamics of a scene and simultaneously output high temporal resolution events and low frame-rate intensity images. In this paper, we propose a single image (potentially blurred) and events based optical flow estimation approach. First, we demonstrate how events can be used to improve flow estimates. To this end, we encode the relation between flow and events effectively by presenting an event-based photometric consistency formulation. Then, we consider the special case of image blur caused by high dynamics in the visual environments and show that including the blur formation in our model further constrains flow estimation. This is in sharp contrast to existing works that ignore the blurred images while our formulation can naturally handle either blurred or sharp images to achieve accurate flow estimation. Finally, we reduce flow estimation, as well as image deblurring, to an alternative optimization problem of an objective function using the primal-dual algorithm. Experimental results on both synthetic and real data (with blurred and non-blurred images) show the superiority of our model in comparison to state-of-the-art approaches.
	16	Proceedings of the IEEE CVPR, 2020, pp. 4919-4928	KFNet: Learning Temporal Camera Relocalization Using Kalman Filtering	Lei Zhou, Zixin Luo, Tianwei Shen, Jiahui Zhang, Mingmin Zhen, Yao Yao, Tian Fang, Long Quan	Hong Kong University of Science and Technology	相关(CNN)	Temporal camera relocalization estimates the pose with respect to each video frame in sequence, as opposed to one-shot relocalization which focuses on a still image. Even though the time dependency has been taken into account, current temporal relocalization methods still generally underperform the state-of-the-art one-shot approaches in terms of accuracy. In this work, we improve the temporal relocalization method by using a network architecture that incorporates Kalman filtering (KFNet) for online camera relocalization. In particular, KFNet extends the scene coordinate regression problem to the time domain in order to recursively establish 2D and 3D correspondences for the pose determination. The network architecture design and the loss formulation are based on Kalman filtering in the context of Bayesian learning. Extensive experiments on multiple relocalization benchmarks demonstrate the high accuracy of KFNet at the top of
	17	Proceedings of the IEEE CVPR, 2020, pp. 11375-11384	Learning Multi-View Camera Relocalization With Graph Neural Networks	Fei Xue, Xin Wu, Shaojun Cai, Junqiu Wang	UISEE Technology Inc.	相关(GNN)	We propose to construct a view graph to excavate the information of the whole given sequence for absolute camera pose estimation. Specifically, we harness GNNs to model the graph, allowing even non-consecutive frames to exchange information with each other. Rather than adopting the regular GNNs directly, we redefine the nodes, edges, and embedded functions to fit the relocalization task. Redesigned GNNs cooperate with CNNs in guiding knowledge propagation and feature extraction respectively to process multi-view high-dimension image features iteratively at different levels. Besides, a general graph-based loss function beyond constraints between consecutive views is employed for training the network in an end-to-end fashion. Extensive experiments conducted on both indoor and outdoor datasets demonstrate that our method outperforms previous approaches especially in large-scale and challenging scenarios.
	18	Proceedings of the IEEE CVPR, 2020, pp. 12144-12153	Camera On-Boarding for Person Re-Identification Using Hypothesis Transfer Learning	Sk Miraj Ahmed, Aske R. Lejbolle, Rameswar Panda, Amit K. Roy-Chowdhury	University of California, Riverside	不相关	Most of the existing approaches for person re-identification consider a static setting where the number of cameras in the network is fixed. An interesting direction, which has received little attention, is to explore the dynamic nature of a camera network, where one tries to adapt the existing re-identification models after on-boarding new cameras, with little additional effort. There have been a few recent methods proposed in person re-identification that attempt to address this problem by assuming the labeled data in the existing network is still available while adding new cameras. This is a strong assumption since there may exist some privacy issues for which one may not have access to those data. Rather, based on the fact that it is easy to store the learned re-identifications models, which mitigates any data privacy concern, we develop an efficient model adaptation approach using hypothesis transfer learning that aims to transfer the knowledge using only source models and limited labeled data, but without using any source camera data from the existing network. Our approach minimizes the effect of negative transfer by finding an optimal weighted combination of multiple source models for transferring the knowledge. Extensive experiments on four challenging benchmark datasets with variable number of cameras well
	19	Proceedings of the IEEE CVPR, 2020, pp. 4968-4978	EventCap: Monocular 3D Capture of High-Speed Human Motions Using an Event Camera	Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, Christian Theobalt	Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, China	相关(CNN)	The high frame rate is a critical requirement for capturing fast human motions. In this setting, existing markerless image-based methods are constrained by the lighting requirement, the high data bandwidth and the consequent high computation overhead. In this paper, we propose EventCap -- the first approach for 3D capturing of high-speed human motions using a single event camera. Our method combines model-based optimization and CNN-based human pose detection to capture high frequency motion details and to reduce the drifting in the tracking. As a result, we can capture fast motions at millisecond resolution with significantly higher data efficiency than using high frame rate videos. Experiments on our new event-based fast human motion dataset demonstrate the effectiveness and accuracy of our method, as well as its robustness to challenging lighting conditions.

	20	Proceedings of the IEEE CVPR, 2020, pp. 13627-13636	End-to-End Camera Calibration for Broadcast Videos	Long Sha, Jennifer Hobbs, Panna Felsen, Xinyu Wei, Patrick Lucey, Sujoy Ganguly	Stats Perform	不相关	The increasing number of vision-based tracking systems deployed in production have necessitated fast, robust camera calibration. In the domain of sport, the majority of current work focuses on sports where lines and intersections are easy to extract, and appearance is relatively consistent across venues. However, for more challenging sports like basketball, those techniques are not sufficient. In this paper, we propose an end-to-end approach for single moving camera calibration across challenging scenarios in sports. Our method contains three key modules: 1) area-based court segmentation, 2) camera pose estimation with embedded templates, 3) homography prediction via a spatial transform network (STN). All three modules are connected, enabling end-to-end training. We evaluate our method on a new college basketball dataset and demonstrate state of the art performance in variable and dynamic environments. We also validate our method on the World Cup 2014 dataset to show its competitive performance against the state-of-the-art methods. Lastly, we show that our method is two orders of magnitude faster than the previous state of the
	21	Proceedings of the IEEE CVPR, 2020, pp. 3320-3329	Learning Event-Based Motion Deblurring	Zhe Jiang, Yu Zhang, Dongqing Zou, Jimmy Ren, Jiancheng Lv, Yebin Liu	SenseTime Research	不相关	Recovering sharp video sequence from a motion-blurred image is highly ill-posed due to the significant loss of motion information in the blurring process. For event-based cameras, however, fast motion can be captured as events at high frame rate, raising new opportunities to exploring effective solutions. In this paper, we start from a sequential formulation of event-based motion deblurring, then show how its optimization can be unfolded with a novel end-toend deep architecture. The proposed architecture is a convolutional recurrent neural network that integrates visual and temporal knowledge of both global and local scales in principled manner. To further improve the reconstruction, we propose a differentiable directional event filtering module to effectively extract rich boundary prior from the evolution of events. We conduct extensive experiments on the synthetic GoPro dataset and a large newly introduced dataset captured by a DAVIS240C camera. The proposed approach achieves state-of-the-art reconstruction quality, and generalizes better to handling real-world motion blur.
	22	Proceedings of the IEEE CVPR, 2020, pp. 6349-6358	Globally Optimal Contrast Maximisation for Event-Based Motion Estimation	Daqi Liu, Alvaro Parra, Tat-Jun Chin	School of Computer Science, The University of Adelaide	不相关	Contrast maximisation estimates the motion captured in an event stream by maximising the sharpness of the motion-compensated event image. To carry out contrast maximisation, many previous works employ iterative optimisation algorithms, such as conjugate gradient, which require good initialisation to avoid converging to bad local minima. To alleviate this weakness, we propose a new globally optimal event-based motion estimation algorithm. Based on branch-and-bound (BnB), our method solves rotational (3DoF) motion estimation on event streams, which supports practical applications such as video stabilisation and attitude estimation. Underpinning our method are novel bounding functions for contrast maximisation, whose theoretical validity is rigorously established. We show concrete examples from public datasets where globally optimal solutions are vital to the success of contrast maximisation. Despite its exact nature, our algorithm is currently able to process a 50,000-event input in approx 300 seconds (a locally optimal solver takes approx 30 seconds on the same input), and has the potential to be further speeded-up using GPUs.
	23	Proceedings of the IEEE CVPR, 2020, pp. 13075-13085	Context R-CNN: Long Term Temporal Context for Per-Camera Object Detection	Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, Jonathan Huang	California Institute of Technology	相关(CNN)	In static monitoring cameras, useful contextual information can stretch far beyond the few seconds typical video understanding models might see: subjects may exhibit similar behavior over multiple days, and background objects remain static. Due to power and storage constraints, sampling frequencies are low, often no faster than one frame per second, and sometimes are irregular due to the use of a motion trigger. In order to perform well in this setting, models must be robust to irregular sampling rates. In this paper we propose a method that leverages temporal context from the unlabeled frames of a novel camera to improve performance at that camera. Specifically, we propose an attention-based approach that allows our model, Context R-CNN, to index into a long term memory bank constructed on a per-camera basis and aggregate contextual features from other frames to boost object detection performance on the current frame. We apply Context R-CNN to two settings: (1) species detection using camera traps, and (2) vehicle detection in traffic cameras, showing in both settings that Context R-CNN leads to performance gains over strong baselines. Moreover, we show that increasing the contextual time horizon leads to improved results. When applied to camera trap data from the Snapshot Serengeti dataset, Context R-CNN with context from up to a month of images outperforms a single-frame baseline by 17.9% mAP, and outperforms S3D (a 3d convolution based baseline) by 11.2% mAP.
	24	Proceedings of the IEEE CVPR, 2020, pp. 5991-6000	Uncertainty Based Camera Model Selection	Michal Polic, Stanislav Steidl, Cenek Albl, Zuzana Kukelova, Tomas Pajdla	CTU in Prague	相关(点云)	The quality and speed of Structure from Motion (SfM) methods depend significantly on the camera model chosen for the reconstruction. In most of the SfM pipelines, the camera model is manually chosen by the user. In this paper, we present a new automatic method for camera model selection in large scale SfM that is based on efficient uncertainty evaluation. We first perform an extensive comparison of classical model selection based on known Information Criteria and show that they do not provide sufficiently accurate results when applied to camera model selection. Then we propose a new Accuracy-based Criterion, which evaluates an efficient approximation of the uncertainty of the estimated parameters in tested models. Using the new criterion, we design a camera model selection method and fine-tune it by machine learning. Our simulated and real experiments demonstrate a significant increase in reconstruction quality as well as a considerable
	25	Proceedings of the IEEE CVPR, 2020, pp. 1701-1710	Event Probability Mask (EPM) and Event Denoising Convolutional Neural Network (EDnCNN) for Neuromorphic Cameras	R. Wes Baldwin, Mohammed Almatrafi, Vijayan Asari, Keigo Hirakawa	Department of Electrical Engineering, University of Dayton	相关(CNN)	This paper presents a novel method for labeling real-world neuromorphic camera sensor data by calculating the likelihood of generating an event at each pixel within a short time window, which we refer to as "event probability mask" or EPM. Its applications include (i) objective benchmarking of event denoising performance, (ii) training convolutional neural networks for noise removal called "event denoising convolutional neural network" (EDnCNN), and (iii) estimating internal neuromorphic camera parameters. We provide the first dataset (DVSNOISE20) of real-world labeled neuromorphic camera events for noise removal.
	1	Proceedings of the IEEE CVPR, 2021, pp. 662-671	Removing Diffraction Image Artifacts in Under-Display Camera via Dynamic Skip Connection Network	Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Chen Change Loy, Jinwei Gu	S-Lab, Nanyang Technological University	不相关	Recent development of Under-Display Camera (UDC) systems provides a true bezel-less and notch-free viewing experience on smartphones (and TV, laptops, tablets), while allowing images to be captured from the selfie camera embedded underneath. In a typical UDC system, the microstructure of the semi-transparent organic light-emitting diode (OLED) pixel array attenuates and diffracts the incident light on the camera, resulting in significant image quality degradation. Oftentimes, noise, flare, haze, and blur can be observed in UDC images. In this work, we aim to analyze and tackle the aforementioned degradation problems. We define a physics-based image formation model to better understand the degradation. In addition, we utilize one of the world's first commodity UDC smartphone prototypes to measure the real-world Point Spread Function (PSF) of the UDC system, and provide a model-based data synthesis pipeline to generate realistically degraded images. We specially design a new domain knowledge-enabled Dynamic Skip Connection Network (DISCNet) to restore the UDC images. We demonstrate the effectiveness of our method through extensive experiments on both synthetic and real UDC data. Our physics-based image formation model and proposed DISCNet can provide foundations for further exploration in UDC image restoration, and even for general diffraction artifact removal in a broader sense.
	2	Proceedings of the IEEE CVPR, 2021, pp. 13274-13283	Camera-Space Hand Mesh Recovery via Semantic Aggregation and Adaptive 2D-1D Registration	Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, Wen Zheng	Y-tech, Kuaishou Technology	不相关	Recent years have witnessed significant progress in 3D hand mesh recovery. Nevertheless, because of the intrinsic 2D-to-3D ambiguity, recovering camera-space 3D information from a single RGB image remains challenging. To tackle this problem, we divide camera-space mesh recovery into two sub-tasks, i.e., root-relative mesh recovery and root recovery. First, joint landmarks and silhouette are extracted from a single input image to provide 2D cues for the 3D tasks. In the root-relative mesh recovery task, we exploit semantic relations among joints to generate a 3D mesh from the extracted 2D cues. Such generated 3D mesh coordinates are expressed relative to a root position, i.e., wrist of the hand. In the root recovery task, the root position is registered to the camera space by aligning the generated 3D mesh back to 2D cues, thereby completing camera-space 3D mesh recovery. Our pipeline is novel in that (1) it explicitly makes use of known semantic relations among joints and (2) it exploits 1D projections of the silhouette and mesh to achieve robust registration. Extensive experiments on popular datasets such as FreiHAND, RHD, and Human3.6M demonstrate that our approach achieves state-of-the-art performance on both root-relative mesh recovery and root recovery. Our code is publicly available at https://github.com/SeanChenxy/HandMesh.

3	Proceedings of the IEEE CVPR, 2021, pp. 5538-5547	EffiScene: Efficient Per-Pixel Rigidity Inference for Unsupervised Joint Learning of Optical Flow, Depth, Camera Pose and Motion Segmentation	Yang Jiao, Trac D. Tran, Guangming Shi	Xidian University	相关	This paper addresses the challenging unsupervised scene flow estimation problem by jointly learning four low-level vision sub-tasks: optical flow F, stereo-depth D, camera pose P and motion segmentation S. Our key insight is that the rigidity of the scene shares the same inherent geometrical structure with object movements and scene depth. Hence, rigidity from S can be inferred by jointly coupling F, D and S to achieve more robust estimation. To this end, we propose a novel scene flow framework named EffiScene with efficient joint rigidity learning, going beyond the existing pipeline with independent auxiliary structures. In EffiScene, we first estimate optical flow and depth at the coarse level and then compute camera pose by Perspective-n-Points method. To jointly learn local rigidity, we design a novel Rigidity From Motion (RfM) layer with three principal components: (i) correlation extraction; (ii) boundary learning; and (iii) outlier exclusion. Final outputs are fused based on the rigid map M_R from RfM at finer levels. To efficiently train EffiScene, two new losses L_bnd and L_unc are designed to prevent trivial solutions and to regularize the flow boundary discontinuity. Extensive experiments on scene flow benchmark KITTI show that our method is effective and significantly improves the state-of-the-art approaches for all sub-tasks, i.e. optical flow (5.19 -> 4.20), depth estimation (3.78 -> 3.46), visual odometry (0.012 -> 0.011) and motion segmentation (0.57 -> 0.62).
4	Proceedings of the IEEE CVPR, 2021, pp. 10204-10212	Pedestrian and Ego-Vehicle Trajectory Prediction From Monocular Camera	Lukas Neumann, Andrea Vedaldi	Visual Recognition Group Faculty of Electrical Engineering Czech Technical University in Prague	相关(GNN)	Predicting future pedestrian trajectory is a crucial component of autonomous driving systems, as recognizing critical situations based only on current pedestrian position may come too late for any meaningful corrective action (e.g. breaking) to take place. In this paper, we propose a new method to predict future position of pedestrians, with respect to a predicted future position of the ego-vehicle, thus giving a assistive/autonomous driving system sufficient time to respond. The method explicitly disentangles actual movement of pedestrians in real world from the ego-motion of the vehicle, using a future pose prediction network trained in self-supervised fashion, which allows the method to observe and predict the intrinsic pedestrian motion in a normalised view, that captures the same real-world location across multiple frames. The method is evaluated on two public datasets, where it achieves state-of-the-art results in pedestrian trajectory prediction from an
5	Proceedings of the IEEE CVPR, 2021, pp. 3446-3455	Back to Event Basics: Self-Supervised Learning of Image Reconstruction for Event Cameras via Photometric Constancy	Federico Paredes-Valles, Guido C. H. E. de Croon	Micro Air Vehicle Laboratory, Delft University of Technology, The Netherlands	不相关	Event cameras are novel vision sensors that sample, in an asynchronous fashion, brightness increments with low latency and high temporal resolution. The resulting streams of events are of high value by themselves, especially for high speed motion estimation. However, a growing body of work has also focused on the reconstruction of intensity frames from the events, as this allows bridging the gap with the existing literature on appearance- and frame-based computer vision. Recent work has mostly approached this problem using neural networks trained with synthetic, ground-truth data. In this work we approach, for the first time, the intensity reconstruction problem from a self-supervised learning perspective. Our method, which leverages the knowledge of the inner workings of event cameras, combines estimated optical flow and the event-based photometric constancy to train neural networks without the need for any ground-truth or synthetic data. Results across multiple datasets show that the performance of the proposed self-supervised approach is in line with the state-of-the-art. Additionally, we propose a novel, lightweight neural network for optical flow estimation that achieves high speed inference with only a minor drop in performance.
6	Proceedings of the IEEE CVPR, 2021, pp. 1397-1406	Depth From Camera Motion and Object Detection	Brent A. Griffin, Jason J. Corso	University of Michigan	相关(CNN)	This paper addresses the problem of learning to estimate the depth of detected objects given some measurement of camera motion (e.g., from robot kinematics or vehicle odometry). We achieve this by 1) designing a recurrent neural network (DBox) that estimates the depth of objects using a generalized representation of bounding boxes and uncalibrated camera movement and 2) introducing the Object Depth via Motion and Detection Dataset (ODMD). ODMD training data are extensible and configurable, and the ODMD benchmark includes 21,600 examples across four validation and test sets. These sets include mobile robot experiments using an end-effector camera to locate objects from the YCB dataset and examples with perturbations added to camera motion or bounding box data. In addition to the ODMD benchmark, we evaluate DBox in other monocular application domains, achieving state-of-the-art results on existing driving and robotics benchmarks and estimating the depth of objects using a camera phone.
7	Proceedings of the IEEE CVPR, 2021, pp. 9959-9968	Learning Neural Representation of Camera Pose with Matrix Representation of Pose Shift via View Synthesis	Yaxuan Zhu, Ruiqi Gao, Siyuan Huang, Song-Chun Zhu, Ying Nian Wu	Department of Statistics, University of California, Los Angeles (UCLA)	相关	How to efficiently represent camera pose is an essential problem in 3D computer vision, especially in tasks like camera pose regression and novel view synthesis. Traditionally, 3D position of the camera is represented by Cartesian coordinate and the orientation is represented by Euler angle or quaternions. These representations are manually designed, which may not be the most efficient representation for downstream tasks. In this work, we propose an approach to learn neural representations of camera poses and 3D scenes, coupled with neural representations of local camera movements. Specifically, the camera pose and 3D scene are represented as vectors and the local camera movement is represented as a matrix operating on the vector of the camera pose. We demonstrate that the camera movement can further be parametrized as a matrix Lie algebra that underlies a rotation system in the neural space. The vector representa-tions are then concatenated and generate the posed 2D image through a decoder network. The model is learned from only posed 2D images and corresponding camera poses, without access to depth or shape. We conduct extensive experiments on synthetic and real datasets. The results show that compared with other camera pose representations, our learned representation is more robust to noise in novel view
8	Proceedings of the IEEE CVPR, 2021, pp. 8436-8444	Positive Sample Propagation Along the Audio-Visual Event Line	Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, Meng Wang	Hefei University of Technology	不相关	Visual and audio signals often coexist in natural environments, forming audio-visual events (AVEs). Given a video, we aim to localize video segments containing an AVE and identify its category. In order to learn discriminative features for a classifier, it is pivotal to identify the helpful (or positive) audio-visual segment pairs while filtering out the irrelevant ones, regardless whether they are synchronized or not. To this end, we propose a new positive sample propagation (PSP) module to discover and exploit the closely related audio-visual pairs by evaluating the relationship within every possible pair. It can be done by constructing an all-pair similarity map between each audio and visual segment, and only aggregating the features from the pairs with high similarity scores. To encourage the network to extract high correlated features for positive samples, a new audio-visual pair similarity loss is proposed. We also propose a new weighting branch to better exploit the temporal correlations in weakly supervised setting. We perform extensive experiments on the public AVE dataset and achieve new state-of-the-art accuracy in both fully and weakly supervised settings, thus verifying the effectiveness of our method.
9	Proceedings of the IEEE CVPR, 2021, pp. 2073-2082	Controllable Image Restoration for Under-Display Camera in Smartphones	Kinam Kwon, Eunhee Kang, Sangwon Lee, Su-Jin Lee, Hyong-Euk Lee, ByungIn Yoo, Jae-Joon Han	Samsung Advanced Institute of Technology (SAIT), South Korea	不相关	Under-display camera (UDC) technology is essential for full-screen display in smartphones and is achieved by removing the concept of drilling holes on display. However, this causes inevitable image degradation in the form of spatially variant blur and noise because of the opaque display in front of the camera. To address spatially variant blur and noise in UDC images, we propose a novel controllable image restoration algorithm utilizing pixel-wise UDC-specific kernel representation and a noise estimator. The kernel representation is derived from an elaborate optical model that reflects the effect of both normal and oblique light incidence. Also, noise-adaptive learning is introduced to control noise levels, which can be utilized to provide optimal results depending on the user preferences. The experiments showed that the proposed method achieved superior quantitative performance as well as higher perceptual quality on both a real-world dataset and a monitor-based aligned dataset compared to conventional image restoration algorithms.
10	Proceedings of the IEEE CVPR, 2021, pp. 4659-4668	Uncertainty-Aware Camera Pose Estimation From Points and Lines	Alexander Vakhitov, Luis Ferraz, Antonio Agudo, Francesc Moreno-Noguer	SLAMCore Ltd., UK	不相关	Perspective-n-Point-and-Line (PnPL) algorithms aim at fast, accurate, and robust camera localization with respect to a 3D model from 2D-3D feature correspondences, being a major part of modern robotic and AR/VR systems. Current point-based pose estimation methods use only 2D feature detection uncertainties, and the line-based methods do not take uncertainties into account. In our setup, both 3D coordinates and 2D projections of the features are considered uncertain. We propose PnP(L) solvers based on EPnP[20] and DLS[14] for the uncertainty-aware pose estimation. We also modify motion-only bundle adjustment to take 3D uncertainties into account. We perform exhaustive synthetic and real experiments on two different visual odometry datasets. The new PnP(L) methods outperform the state-of-the-art on real data in isolation, showing an increase in mean translation accuracy by 18% on a representative subset of KITTI, while the new uncertain refinement improves pose accuracy for most of the solvers, e.g. decreasing mean translation error for the EPnP by 16% compared to the standard refinement on the same dataset. The code is available at https://alexandervakhitov.github.io/uncertain-pnp/.

11	Proceedings of the IEEE CVPR, 2021, pp. 32-42	Fusing the Old with the New: Learning Relative Camera Pose with Geometry-Guided Uncertainty	Bingbing Zhuang, Manmohan Chandraker	NEC Labs America	相关(DNN)	Learning methods for relative camera pose estimation have been developed largely in isolation from classical geometric approaches. The question of how to integrate predictions from deep neural networks (DNNs) and solutions from geometric solvers, such as the 5-point algorithm, has as yet remained under-explored. In this paper, we present a novel framework that involves probabilistic fusion between the two families of predictions during network training, with a view to leveraging their complementary benefits in a learnable way. The fusion is achieved by learning the DNN uncertainty under explicit guidance by the geometric uncertainty, thereby learning to take into account the geometric solution in relation to the DNN prediction. Our network features a self-attention graph neural network, which drives the learning by enforcing strong interactions between different correspondences and potentially modeling complex relationships between points. We propose motion parmeterizations suitable for learning and show that our method achieves state-of-the-art performance on the challenging DeMoN and ScanNet datasets. While we focus on relative pose, we envision that our pipeline is broadly applicable for fusing classical geometry and deep learning.
12	Proceedings of the IEEE CVPR, 2021, pp. 6297-6307	End-to-End High Dynamic Range Camera Pipeline Optimization	Nicolas Robidoux, Luis E. Garcia Capel, Dong-eun Seo, Avinash Sharma, Federico Ariza, Felix Heide	Algolux	相关(CNN)	With a 280 dB dynamic range, the real world is a High Dynamic Range (HDR) world. Today's sensors cannot record this dynamic range in a single shot. Instead, HDR cameras acquire multiple measurements with different exposures, gains and photodiodes, from which an Image Signal Processor (ISP) reconstructs an HDR image. HDR image recovery for dynamic scenes is an open challenge because of motion and because stitched captures have different noise characteristics, resulting in artefacts that the ISP has to resolve---in real time and at triple-digit megapixel resolutions. Traditionally, hardware ISP settings used by downstream vision modules have been chosen by domain experts. Such frozen camera designs are then used for training data acquisition and supervised learning of downstream vision modules. We depart from this paradigm and formulate HDR ISP hyperparameter search as an end-to-end optimization problem. We propose a mixed 0th and 1st-order block coordinate descent optimizer to jointly learn ISP and detector network weights using RAW image data augmented with emulated SNR transition region artefacts. We assess the proposed method for human vision and image understanding. For automotive object detection, the method improves mAP and mAR by 33% compared to expert-tuning and by 22% compared to recent state-of-the-art. The method is validated in an HDR laboratory rig and in the field, outperforming conventional handcrafted HDR imaging and vision pipelines in all
13	Proceedings of the IEEE CVPR, 2021, pp. 13784-13793	DyGLIP: A Dynamic Graph Model With Link Prediction for Accurate Multi-Camera Multiple Object Tracking	Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, Khoa Luu	Concordia University, CANADA	不相关	Multi-Camera Multiple Object Tracking (MC-MOT) is a significant computer vision problem due to its emerging applicability in several real-world applications. Despite a large number of existing works, solving the data association problem in any MC-MOT pipeline is arguably one of the most challenging tasks. Developing a robust MC-MOT system, however, is still highly challenging due to many practical issues such as inconsistent lighting conditions, varying object movement patterns, or the trajectory occlusions of the objects between the cameras. To address these problems, this work, therefore, proposes a new Dynamic Graph Model with Link Prediction (DyGLIP) approach to solve the data association task. Compared to existing methods, our new model offers several advantages, including better feature representations and the ability to recover from lost tracks during camera transitions. Moreover, our model works gracefully regardless of the overlapping ratios between the cameras. Experimental results show that we outperform existing MC-MOT algorithms by a large margin on several practical datasets. Notably, our model works favorably on online settings but can be extended to an incremental approach for large-scale datasets.
14	Proceedings of the IEEE CVPR, 2021, pp. 414-423	Neural Reprojection Error: Merging Feature Learning and Camera Pose Estimation	Hugo Germain, Vincent Lepetit, Guillaume Bourmaud	LIGM, ´Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-vall´ee, France	不相关	Absolute camera pose estimation is usually addressed by sequentially solving two distinct subproblems: First a feature matching problem that seeks to establish putative 2D-3D corresponden-ces, and then a Perspective-n-Point problem that minimizes, w.r.t. the camera pose, the sum of so-called Reprojection Errors (RE). We argue that generating putative 2D-3D corresponden-ces 1) leads to an important loss of information that needs to be compensated as far as possible, within RE, through the choice of a robust loss and the tuning of its hyperparameters and 2) may lead to an RE that conveys erroneous data to the pose estimator. In this paper, we introduce the Neural Reprojection Error (NRE) as a substitute for RE. NRE allows to rethink the camera pose estimation problem by merging it with the feature learning problem, hence leveraging richer information than 2D-3D correspondences and eliminating the need for choosing a robust loss and its hyperparameters. Thus NRE can be used as training loss to learn image descriptors tailored for pose estimation. We also propose a coarse-to-fine optimization method able to very efficiently minimize a sum of NRE terms w.r.t. the camera pose. We experimentally demonstrate that NRE is a good substitute for RE as it significantly improves both the robustness and the accuracy of the camera pose estimate while being computationally and memory highly efficient. From a broader point of view, we believe this new way of merging deep learning and 3D geometry may be useful
15	Proceedings of the IEEE CVPR, 2021, pp. 3247-3257	Back to the Feature: Learning Robust Camera Localization From Pixels To Pose	Paul-Edouard Sarlin, Ajaykumar Unagar, Mans Larsson, Hugo Germain, Carl Toft, Viktor Larsson, Marc Pollefeys, Vincent Lepetit, Lars Hammarstrand, Fredrik Kahl, Torsten Sattler	Department of Computer Science, ETH Zurich	不相关	Camera pose estimation in known scenes is a 3D geometry task recently tackled by multiple learning algorithms. Many regress precise geometric quantities, like poses or 3D points, from an input image. This either fails to generalize to new viewpoints or ties the model parameters to a specific scene. In this paper, we go Back to the Feature: we argue that deep networks should focus on learning robust and invariant visual features, while the geometric estimation should be left to principled algorithms. We introduce PixLoc, a scene-agnostic neural network that estimates an accurate 6-DoF pose from an image and a 3D model. Our approach is based on the direct alignment of multiscale deep features, casting camera localization as metric learning. PixLoc learns strong data priors by end-to-end training from pixels to pose and exhibits exceptional generalization to new scenes by separating model parameters and scene geometry. The system can localize in large environments given coarse pose priors but also improve the accuracy of sparse feature matching by jointly refining keypoints and poses with little overhead. The code will be publicly available at github.com/cvg/pixloc.
16	Proceedings of the IEEE CVPR, 2021, pp. 1831-1841	Learning Camera Localization via Dense Scene Matching	Shitao Tang, Chengzhou Tang, Rui Huang, Siyu Zhu, Ping Tan	Simon Fraser University	相关(CNN)	Camera localization aims to estimate 6 DoF camera poses from RGB images. Traditional methods detect and match interest points between a query image and a pre-built 3D model. Recent learning-based approaches encode scene structures into a specific convolutional neural network(CNN) and thus are able to predict dense coordinates from RGB images. However, most of them require re-training or re-adaption for a new scene and have difficulties in handling large-scale scenes due to limited network capacity. We present a new method for scene agnostic camera localization using dense scene matching (DSM), where the cost volume is constructed between a query image and a scene. The cost volume and the corresponding coordinates are processed by a CNN to predict dense coordinates. Camera poses can then be solved by PnP algorithms. In addition, our method can be extended to temporal domain, giving extra performan-ce boost during testing time. Our scene-agnostic approach achieves comparable accuracy as the existing scene-specific approaches on the 7scenes and Cambridge benchmark. This approach also remarkably outperforms state-of-the-art scene-agnostic dense coordinate regression network SANet.
17	Proceedings of the IEEE CVPR, 2021, pp. 15638-15647	Event-Based Bispectral Photometry Using Temporally Modulated Illumination	Tsuyoshi Takatani, Yuzuha Ito, Ayaka Ebisu, Yinqiang Zheng, Takahito Aoto	University of Tsukuba, Japan	相关	Analysis of bispectral difference plays a critical role in various applications that involve rays propagating in a light absorbing medium. In general, the bispectral difference is obtained by subtracting signals at two individual wavelengths captured by ordinary digital cameras, which tends to inherit the drawbacks of conventional cameras in dynamic range, response speed and quantization precision. In this paper, we propose a novel method to obtain a bispectral difference image using an event camera with temporally modulated illumination. Our method is rooted in a key observation on the analogy between the bispectral photometry principle of the participating medium and the event generating mechanism in an event camera. By carefully modulating the bispectral illumination, our method allows to read out the bispectral difference directly from triggered events. Experiments using a prototype imaging system have verified the feasibility of this novel usage of event cameras in photometry based vision tasks, such as 3D shape reconstruction in water.

2021	18	Proceedings of the IEEE CVPR, 2021, pp. 4855-4864	Joint Noise-Tolerant Learning and Meta Camera Shift Adaptation for Unsupervised Person Re-Identification	Fengxiang Yang, Zhun Zhong, Zhiming Luo, Yuanzheng Cai, Yaojin Lin, Shaozi Li, Nicu Sebe	Department of Artificial Intelligence, School of Informatics, Xiamen University	不相关	This paper considers the problem of unsupervised person re-identification (re-ID), which aims to learn discriminative models with unlabeled data. One popular method is to obtain pseudo-label by clustering and use them to optimize the model. Although this kind of approach has shown promising accuracy, it is hampered by 1) noisy labels produced by clustering and 2) feature variations caused by camera shift. The former will lead to incorrect optimization and thus hinders the model accuracy. The latter will result in assigning the intra-class samples of different cameras to different pseudo-label, making the model sensitive to camera variations. In this paper, we propose a unified framework to solve both problems. Concretely, we propose a Dynamic and Symmetric Cross-Entropy loss (DSCE) to deal with noisy samples and a camera-aware meta-learning algorithm (MetaCam) to adapt camera shift. DSCE can alleviate the negative effects of noisy samples and accommodate the change of clusters after each clustering step. MetaCam simulates cross-camera constraint by splitting the training data into meta-train and meta-test based on camera IDs. With the interacted gradient from meta-train and meta-test, the model is enforced to learn camera-invariant features. Extensive experiments on three re-ID benchmarks show the effectiveness and the complementary of the proposed DSCE and MetaCam. Our method outperforms the state-of-the-art methods on both fully unsupervised re-ID and unsupervised domain adaptive re-ID.
	19	Proceedings of the IEEE CVPR, 2021, pp. 7535-7545	Mesoscopic Photogrammetry With an Unstabilized Phone Camera	Kevin C. Zhou, Colin Cooke, Jaehee Park, Ruobing Qian, Roarke Horstmeyer, Joseph A. Izatt, Sina Farsiu	Duke University, Durham, NC	相关(CNN)	We present a feature-free photogrammetric technique that enables quantitative 3D mesoscopic (mm-scale height variation) imaging with tens-of-micron accuracy from sequences of images acquired by a smartphone at close range (several cm) under freehand motion without additional hardware. Our end-to-end, pixel-intensity-based approach jointly registers and stitches all the images by estimating a coaligned height map, which acts as a pixel-wise radial deformation field that orthorectifies each camera image to allow plane-plus-parallax registration. The height maps themselves are reparameterized as the output of an untrained encoder-decoder convolutional neural network (CNN) with the raw camera images as the input, which effectively removes many reconstruction artifacts. Our method also jointly estimates both the camera's dynamic 6D pose and its distortion using a nonparametric model, the latter of which is especially important in mesoscopic applications when using cameras not designed for imaging at short working distances, such as smartphone cameras. We also propose strategies for reducing computation time and memory, applicable to other multi-frame registration problems. Finally, we demonstrate our method using sequences of multi-megapixel images captured by an unstabilized smartphone on a
	20	Proceedings of the IEEE CVPR, 2021, pp. 13134-13143	Wide-Baseline Multi-Camera Calibration Using Person Re-Identification	Yan Xu, Yu-Jhe Li, Xinshuo Weng, Kris Kitani	Carnegie Mellon University	不相关	We address the problem of estimating the 3D pose of a network of cameras for large-environment wide-baseline scenarios, e.g., cameras for construction sites, sports stadiums, and public spaces. This task is challenging since detecting and matching the same 3D keypoint observed from two very different camera views is difficult, making standard structure-from-motion (SfM) pipelines inapplicable. In such circumstances, treating people in the scene as "keypoints" and associating them across different camera views can be an alternative method for obtaining correspondences. Based on this intuition, we propose a method that uses ideas from person re-identification (re-ID) for wide-baseline camera calibration. Our method first employs a re-ID method to associate human bounding boxes across cameras, then converts bounding box correspondences to point correspondences, and finally solves for camera pose using multi-view geometry and bundle adjustment. Since our method does not require specialized calibration targets except for visible people, it applies to situations where frequent calibration updates are required. We perform extensive experiments on datasets captured from scenes of different sizes, camera settings (indoor and outdoor), and human activities (walking, playing basketball, construction). Experiment results show that our method achieves similar performance to standard SfM methods relying on manually labeled point correspondences.
	21	Proceedings of the IEEE CVPR, 2021, pp. 12596-12606	Multi-Shot Temporal Event Localization: A Benchmark	Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, Philip H. S. Torr	Huazhong University of Science and Technology	不相关	Current developments in temporal event or action localization usually target actions captured by a single camera. However, extensive events or actions in the wild may be captured as a sequence of shots by multiple cameras at different positions. In this paper, we propose a new and challenging task called multi-shot temporal event localization, and accordingly, collect a large-scale dataset called MUlti-Shot EventS (MUSES). MUSES has 31,477 event instances for a total of 716 video hours. The core nature of MUSES is the frequent shot cuts, for an average of 19 shots per instance and 176 shots per video, which induces large intra-instance variations. Our comprehensive evaluations show that the state-of-the-art method in temporal action localization only achieves an mAP of 13.1% at IoU=0.5. As a minor contribution, we present a simple baseline approach for handling the intra-instance variations, which reports an mAP of 18.9% on MUSES and 56.9% on THUMOS14 at IoU=0.5. To facilitate research in this direction, we release the dataset and the project code at https://songbai.site/muses/.
	22	Proceedings of the IEEE CVPR, 2021, pp. 4937-4946	Spatiotemporal Registration for Event-Based Visual Odometry	Daqi Liu, Alvaro Parra, Tat-Jun Chin	School of Computer Science, The University of Adelaide	不相关	A useful application of event sensing is visual odometry, especially in settings that require high-temporal resolution. The state-of-the-art method of contrast maximisation recovers the motion from a batch of events by maximising the contrast of the image of warped events. However, the cost scales with image resolution and the temporal resolution can be limited by the need for large batch sizes to yield sufficient structure in the contrast image (see supplementary material for demonstration program). In this work, we propose spatiotemporal registration as a compelling technique for event-based rotational motion estimation. We theoretically justify the approach and establish its fundamental and practical advantages over contrast maximisation. In particular, spatiotemporal registration also produces feature tracks as a by-product, which directly supports an efficient visual odometry pipeline with graph-based optimisation for motion averaging. The simplicity of our visual odometry pipeline allows it to process more than 1 M events/second. We also contribute a new event dataset for visual odometry, where motion sequences with large velocity variations were acquired using a high-precision robot arm. Our dataset will be published after the reviewing period.
	23	Proceedings of the IEEE CVPR, 2021, pp. 7772-7781	Turning Frequency to Resolution: Video Super-Resolution via Event Cameras	Yongcheng Jing, Yiding Yang, Xinchao Wang, Mingli Song, Dacheng Tao	The University of Sydney	相关	State-of-the-art video super-resolution (VSR) methods focus on exploiting inter- and intra-frame correlations to estimate high-resolution (HR) video frames from low-resolution (LR) ones. In this paper, we study VSR from an exotic perspective, by explicitly looking into the role of temporal frequency of video frames. Through experiments, we observe that a higher frequency, and hence a smaller pixel displacement between consecutive frames, tends to deliver favorable super-resolved results. This discovery motivates us to introduce Event Cameras, a novel sensing device that responds instantly to pixel intensity changes and produces up to millions of asynchronous events per second, to facilitate VSR. To this end, we propose an Event-based VSR framework (E-VSR), of which the key component is an asynchronous interpolation (EAI) module that reconstructs a high-frequency (HF) video stream with uniform and tiny pixel displacements between neighboring frames from an event stream. The derived HF video stream is then encoded into a VSR module to recover the desired HR videos. Furthermore, an LR bi-directional interpolation loss and an HR self-supervision loss are also introduced to respectively regulate the EAI and VSR modules. Experiments on both real-world and synthetic datasets demonstrate that the proposed approach yields results superior to the state of the art.
	24	Proceedings of the IEEE CVPR, 2021, pp. 7700-7709	Neural Camera Simulators	Hao Ouyang, Zifan Shi, Chenyang Lei, Ka Lung Law, Qifeng Chen	HKUST	相关	We present a controllable camera simulator based on deep neural networks to synthesize raw image data under different camera settings, including exposure time, ISO, and aperture. The proposed simulator includes an exposure module that utilizes the principle of modern lens designs for correcting the luminance level. It also contains a noise module using the noise level function and an aperture module with adaptive attention to simulate the side effects on noise and defocus blur. To facilitate the learning of a simulator model, we collect a dataset of the 10,000 raw images of 450 scenes with different exposure settings. Quantitative experiments and qualitative comparisons show that our approach outperforms relevant baselines in raw data synthesize on multiple cameras. Furthermore, the camera simulator enables various applications, including large-aperture enhancement, HDR, auto exposure, and data augmentation for training local feature detectors. Our work represents the first attempt to simulate a camera sensor's behavior leveraging both the advantage of traditional raw sensor features and the power of data-driven deep learning.

25	Proceedings of the IEEE CVPR, 2021, pp. 3258-3268	Wide-Baseline Relative Camera Pose Estimation With Directional Learning	Kefan Chen, Noah Snavely, Ameesh Makadia	Google Research	不相关	Modern deep learning techniques that regress the relative camera pose between two images have difficulty dealing with challenging scenarios, such as large camera motions resulting in occlusions and significant changes in perspective that leave little overlap between images. These models continue to struggle even with the benefit of large supervised training datasets. To address the limitations of these models, we take inspiration from techniques that show regressing keypoint locations in 2D and 3D can be improved by estimating a discrete distribution over keypoint locations. Analogously, in this paper we explore improving camera pose regression by instead predicting a discrete distribution over camera poses. To realize this idea, we introduce DirectionNet, which estimates discrete distributions over the 5D relative pose space using a novel parameterization to make the estimation problem tractable. Specifically, DirectionNet factorizes relative camera pose, specified by a 3D rotation and a translation direction, into a set of 3D direction vectors. Since 3D directions can be identified with points on the sphere, DirectionNet estimates discrete distributions on the sphere as its output. We evaluate our model on challenging synthetic and real pose estimation datasets constructed from Matterport3D and InteriorNet. Promising results show a near 50% reduction in error over direct regression methods.
26	Proceedings of the IEEE CVPR, 2021, pp. 11926-11935	Intra-Inter Camera Similarity for Unsupervised Person Re-Identification	Shiyu Xuan, Shiliang Zhang	Department of Computer Science, School of EECS, Peking University	相关(CNN)	Most of unsupervised person Re-Identification (Re-ID) works produce pseudo-labels by measuring the feature similarity without considering the distribution discrepancy among cameras, leading to degraded accuracy in label computation across cameras. This paper targets to address this challenge by studying a novel intra-inter camera similarity for pseudo-label generation. We decompose the sample similarity computation into two stage, i.e., the intra-camera and inter-camera computations, respectively. The intra-camera computation directly leverages the CNN features for similarity computation within each camera. Pseudo-labels generated on different cameras train the re-id model in a multi-branch network. The second stage considers the classification scores of each sample on different cameras as a new feature vector. This new feature effectively alleviates the distribution discrepancy among cameras and generates more reliable pseudo-labels. We hence train our re-id model in two stages with intra-camera and inter-camera pseudo-labels, respectively. This simple intra-inter camera similarity produces surprisingly good performance on multiple datasets, e.g., achieves rank-1 accuracy of 89.5% on the Market1501 dataset, outperforming the recent unsupervised works by 9+%, and is comparable with the latest transfer learning
27	Proceedings of the IEEE CVPR, 2021, pp. 8425-8435	Towards Bridging Event Captioner and Sentence Localizer for Weakly Supervised Dense Event Captioning	Shaoxiang Chen, Yu-Gang Jiang	Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University	不相关	Dense Event Captioning (DEC) aims to jointly localize and describe multiple events of interest in untrimmed videos, which is an advancement of the conventional video captioning task (generating a single sentence description for a trimmed video). Weakly Supervised Dense Event Captioning (WS-DEC) goes one step further by not relying on human-annotated temporal event boundaries. However, there are few methods trying to tackle this task, and how to connect localization and description remains an open problem. In this paper, we demonstrate that under weak supervision, the event captioning module and localization module should be more closely bridged in order to improve description performance. Different from previous approach-es, in our method, the event captioner generates a sentence from a video segment and feeds it to the sentence localizer to reconstruct the segment, and the localizer produces word importa-nce weights as a guidance for the captioner to improve event description. To further bridge the sentence localizer and event captioner, a concept learner is adopted as the basis of the senten-ce localizer, which can be utilized to construct an induced set of concept features to enhance video features and improve the event captioner. Finally, our proposed method outperforms state-of-the-art WS-DEC methods on the ActivityNet Captions dataset.
28	Proceedings of the IEEE CVPR, 2021, pp. 14235-14244	Event-Based Synthetic Aperture Imaging With a Hybrid Network	Xiang Zhang, Wei Liao, Lei Yu, Wen Yang, Gui-Song Xia	Wuhan University	相关(CNN)	Synthetic aperture imaging (SAI) is able to achieve the see through effect by blurring out the off-focus foreground occlusions and reconstructing the in-focus occluded targets from multi-view images. However, very dense occlusions and extreme lighting conditions may bring significant disturbances to the SAI based on conventional frame-based cameras, leading to perfor-mance degeneration. To address these problems, we propose a novel SAI system based on the event camera which can produce asynchronous events with extremely low latency and high dynamic range. Thus, it can eliminate the interference of dense occlusions by measuring with almost continuous views, and simultaneously tackle the over/under exposure problems. To reconstruct the occluded targets, we propose a hybrid encoder-decoder network composed of spiking neural networks (SNNs) and convolutional neural networks (CNNs). In the hybrid network, the spatio-temporal information of the collected events is first encoded by SNN layers, and then transformed to the visual image of the occluded targets by a style-transfer CNN decoder. Through experiments, the proposed method shows remarkable performance in dealing with very dense occlusions and extreme lighting conditions, and high quality visual images can be reconstructed using pure event data.
29	Proceedings of the IEEE CVPR, 2021, pp. 15759-15768	Camera Pose Matters: Improving Depth Prediction by Mitigating Pose Distribution Bias	Yunhan Zhao, Shu Kong, Charless Fowlkes	UC Irvine	相关(CNN)	Monocular depth predictors are typically trained on large-scale training sets which are naturally biased w.r.t the distribution of camera poses. As a result, trained predictors fail to make reliable depth predictions for testing examples captured under uncommon camera poses. To address this issue, we propose two novel techniques that exploit the camera pose during training and prediction. First, we introduce a simple perspective-aware data augmentation that synthesizes new training examples with more diverse views by perturbing the existing ones in a geometrically consistent manner. Second, we propose a conditional model that exploits the per-image camera pose as prior knowledge by encoding it as a part of the input. We show that jointly applying the two methods improves depth prediction on images captured under uncommon and even never-before-seen camera poses. We show that our methods improve performan-ce when applied to a range of different predictor architectures. Lastly, we show that explicitly encoding the camera pose distribution improves the generalization performance of a syntheti-cally trained depth predictor
30	Proceedings of the IEEE CVPR, 2021, pp. 12507-12516	Radar-Camera Pixel Depth Association for Depth Completion	Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, Praveen Narayanan	Michigan State University	相关(LiDAR)	While radar and video data can be readily fused at the detection level, fusing them at the pixel level is potentially more beneficial. This is also more challenging in part due to the sparsity of radar, but also because automotive radar beams are much wider than a typical pixel combined with a large baseline between camera and radar, which results in poor association between radar pixels and color pixel. A consequence is that depth completion methods designed for LiDAR and video fare poorly for radar and video. Here we propose a radar-to-pixel association stage which learns a mapping from radar returns to pixels. This mapping also serves to densify radar returns. Using this as a first stage, followed by a more traditional depth completion method, we are able to achieve image-guided depth completion with radar and video. We demonstrate performance superior to camera and radar alone on the nuScenes dataset. Our source code is available at https://github.com/longyunf/rc-pda.
31	Proceedings of the IEEE CVPR, 2021, pp. 9179-9188	Image Restoration for Under-Display Camera	Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, Timothy Large	IFP, UIUC	不相关	The new trend of full-screen devices encourages us to position a camera behind a screen. Removing the bezel and centralizing the camera under the screen brings larger display-to-body ratio and enhances eye contact in video chat, but also causes image degradation. In this paper, we focus on a newly-defined Under-Display Camera (UDC), as a novel real-world single image restoration problem. First, we take a 4k Transparent OLED (T-OLED) and a phone Pentile OLED (P-OLED) and analyze their optical systems to understand the degradation. Second, we design a Monitor-Camera Imaging System (MCIS) for easier real pair data acquisition, and a model-based data synthesizing pipeline to generate Point Spread Function (PSF) and UDC data only from display pattern and camera measurements. Finally, we resolve the complicated degradation using deconvolution-based pipeline and learning-based methods. Our model demonstrates a real-time high-quality restoration. The presented methods and results reveal the promising research values and directions of UDC.
32	Proceedings of the IEEE CVPR, 2021, pp. 14760-14770	Indoor Lighting Estimation Using an Event Camera	Zehao Chen, Qian Zheng, Peisong Niu, Huajin Tang, Gang Pan	College of Computer Science and Technology, Zhejiang University, Hangzhou, China	不相关	Image-based methods for indoor lighting estimation suffer from the problem of intensity-distance ambiguity. This paper introduces a novel setup to help alleviate the ambiguity based on the event camera. We further demonstrate that estimating the distance of a light source becomes a well-posed problem under this setup, based on which an optimization-based method and a learning-based method are proposed. Our experimental results validate that our approaches not only achieve superior performance for indoor lighting estimation (especially for the close light) but also significantly alleviate the intensity-distance ambiguity.

	33	Proceedings of the IEEE CVPR, 2021, pp. 2463-2473	Feature-Level Collaboration: Joint Unsupervised Learning of Optical Flow, Stereo Depth and Camera Motion	Cheng Chi, Qingjie Wang, Tianyu Hao, Peng Guo, Xin Yang	Huazhong University of Science and Technology	相关(CNN)	Precise estimation of optical flow, stereo depth and camera motion are important for the real-world 3D scene understanding and visual perception. Since the three tasks are tightly coupled with the inherent 3D geometric constraints, current studies have demonstrated that the three tasks can be improved through jointly optimizing geometric loss functions of several individual networks. In this paper, we show that effective feature-level collaboration of the networks for the three respective tasks could achieve much greater performance improvement for all three tasks than only loss-level joint optimization. Specifically, we propose a single network to combine and improve the three tasks. The network extracts the features of two consecutive stereo images, and simultaneously estimates optical flow, stereo depth and camera motion. The whole network mainly contains four parts: (I) a feature-sharing encoder to extract features of input images, which can enhance features' representation ability; (II) a pooled decoder to estimate both optical flow and stereo depth; (III) a camera pose estimation module which fuses optical flow and stereo depth information; (IV) a cost volume complement module to improve the performance of optical flow in static and occluded regions. Our method achieves state-of-the-art performance among the joint unsupervised methods, including optical flow and stereo depth estimation on KITTI 2012 and 2015 benchmarks, and camera motion estimation on KITTI VO dataset.
	34	Proceedings of the IEEE CVPR, 2021, pp. 8544-8554	Robust Neural Routing Through Space Partitions for Camera Relocalization in Dynamic Indoor Environments	Siyan Dong, Qingnan Fan, He Wang, Ji Shi, Li Yi, Thomas Funkhouser, Baoquan Chen, Leonidas J. Guibas	Shandong University	不相关	Localizing the camera in a known indoor environment is a key building block for scene mapping, robot navigation, AR, etc. Recent advances estimate the camera pose via optimization over the 2D/3D-3D correspondences established between the coordinates in 2D/3D camera space and 3D world space. Such a mapping is estimated with either a convolution neural network or a decision tree using only the static input image sequence, which makes these approaches vulnerable to dynamic indoor environments that are quite common yet challenging in the real world. To address the aforementioned issues, in this paper, we propose a novel outlier-aware neural tree which bridges the two worlds, deep learning and decision tree approaches. It builds on three important blocks: (a) a hierarchical space partition over the indoor scene to construct the decision tree; (b) a neural routing function, implemented as a deep classification network, employed for better 3D scene understanding; and (c) an outlier rejection module used to filter out dynamic points during the hierarchical routing process. Our proposed algorithm is evaluated on the RIO-10 benchmark developed for camera relocalization in dynamic indoor environments. It achieves robust neural routing through space partitions and outperforms the state-of-the-art approaches by around 30% on
	35	Proceedings of the IEEE CVPR, 2021, pp. 16155-16164	Time Lens: Event-Based Video Frame Interpolation	Stepan Tulyakov, Daniel Gehrig, Stamatios Georgoulis, Julius Erbach, Mathias Gehrig, Yuanyou Li, Davide Scaramuzza	Huawei Technologies, Zurich Research Center	不相关	State-of-the-art frame interpolation methods generate intermediate frames by inferring object motions in the image from consecutive key-frames. In the absence of additional information, first-order approximations, i.e. optical flow, must be used, but this choice restricts the types of motions that can be modeled, leading to errors in highly dynamic scenarios. Event cameras are novel sensors that address this limitation by providing auxiliary visual information in the blind-time between frames. They asynchronously measure per-pixel brightness changes and do this with high temporal resolution and low latency. Event-based frame interpolation methods typically adopt a synthesis-based approach, where predicted frame residuals are directly applied to the key-frames. However, while these approaches can capture non-linear motions they suffer from ghosting and perform poorly in low-texture regions with few events. Thus, synthesis-based and flow-based approaches are complementary. In this work, we introduce Time Lens, a novel method that leverages the advantages of both. We extensively evaluate our method on three synthetic and two real benchmarks where we show an up to 5.21 dB improvement in terms of PSNR over state-of-the-art frame-based and event-based methods. Finally, we release a new large-scale dataset in highly
	36	Proceedings of the IEEE CVPR, 2021, pp. 6112-6122	MonoRec: Semi-Supervised Dense Reconstruction in Dynamic Environments From a Single Moving Camera	Felix Wimbauer, Nan Yang, Lukas von Stumberg, Niclas Zeller, Daniel Cremers	Technical University of Munich	相关(CNN)	In this paper, we propose MonoRec, a semi-supervised monocular dense reconstruction architecture that predicts depth maps from a single moving camera in dynamic environments. MonoRec is based on a multi-view stereo setting which encodes the information of multiple consecutive images in a cost volume. To deal with dynamic objects in the scene, we introduce a MaskModule that predicts moving object masks by leveraging the photometric inconsistencies encoded in the cost volumes. Unlike other multi-view stereo methods, MonoRec is able to reconstruct both static and moving objects by leveraging the predicted masks. Furthermore, we present a novel multi-stage training scheme with a semi-supervised loss formulation that does not require LiDAR depth values. We carefully evaluate MonoRec on the KITTI dataset and show that it achieves state-of-the-art performance compared to both multi-view and single-view methods. With the model trained on KITTI, we further demonstrate that MonoRec is able to generalize well to both the Oxford RobotCar dataset and the more challenging TUM-Mono dataset recorded by a handheld camera. Code and related materials are available at https://vision.in.tum.de/research/monorec.
	1	Proceedings of the IEEE CVPR, 2022, pp. 6845-6854	DiffPoseNet: Direct Differentiable Camera Pose Estimation	Chethan M. Parameshwara, Gokul Hari, Cornelia Fermüller, Nitin J. Sanket, Yiannis Aloimonos	University of Maryland, College Park College Park, MD	不相关	Current deep neural network approaches for camera pose estimation rely on scene structure for 3D motion estimation, but this decreases the robustness and thereby makes cross-dataset generalization difficult. In contrast, classical approaches to structure from motion estimate 3D motion utilizing optical flow and then compute depth. Their accuracy, however, depends strongly on the quality of the optical flow. To avoid this issue, direct methods have been proposed, which separate 3D motion from depth estimation but compute 3D motion using only image gradients in the form of normal flow. In this paper, we introduce a network NFlowNet, for normal flow estimation which is used to enforce robust and direct constraints. In particular, normal flow is used to estimate relative camera pose based on the cheirality (depth positivity) constraint. We achieve this by formulating the optimization problem as a differentiable cheirality layer, which allows for end-to-end learning of camera pose. We perform extensive qualitative and quantitative evaluation of the proposed DiffPoseNet's sensitivity to noise and its generalization across datasets. We compare our approach to existing state-of-the-art methods on KITTI, TartanAir, and TUM-RGBD datasets
	2	Proceedings of the IEEE CVPR, 2022, pp. 17632-17641	Noise2NoiseFlow: Realistic Camera Noise Modeling Without Clean Images	Ali Maleky, Shayan Kousha, Michael S. Brown, Marcus A. Brubaker	York University	不相关	Image noise modeling is a long-standing problem with many applications in computer vision. Early attempts that propose simple models, such as signal-independent additive white Gaussian noise or the heteroscedastic Gaussian noise model (a.k.a., camera noise level function) are not sufficient to learn the complex behavior of the camera sensor noise. Recently, more complex learning-based models have been proposed that yield better results in noise synthesis and downstream tasks, such as denoising. However, their dependence on supervised data (i.e., paired clean images) is a limiting factor given the challenges in producing ground-truth images. This paper proposes a framework for training a noise model and a denoiser simultaneously while relying only on pairs of noisy images rather than noisy/clean paired image data. We apply this framework to the training of the Noise Flow architecture. The noise synthesis and density estimation results show that our framework outperforms previous signal-processing-based noise models and is on par with its supervised counterpart. The trained denoiser is also shown to significantly improve upon both supervised and weakly supervised baseline denoising approaches. The results indicate that the joint training of a denoiser and a noise model yields significant improvements in the denoiser.
	3	Proceedings of the IEEE CVPR, 2022, pp. 8259-8268	SceneSqueezer: Learning To Compress Scene for Camera Relocalization	Luwei Yang, Rakesh Shrestha, Wenbo Li, Shuaicheng Liu, Guofeng Zhang, Zhaopeng Cui, Ping Tan	Simon Fraser University	不相关	Standard visual localization methods build a priori 3D model of a scene which is used to establish correspondences against the 2D keypoints in a query image. Storing these pre-built 3D scene models can be prohibitively expensive for large-scale environments, especially on mobile devices with limited storage and communication bandwidth. We design a novel framework that compresses a scene while still maintaining localization accuracy. The scene is compressed in three stages: first, the database frames are clustered using pairwise co-visibility information. Then, a learned point selection module prunes the points in each cluster taking into account the final pose estimation accuracy. In the final stage, the features of the selected points are further compressed using learned quantization. Query image registration is done using only the compressed scene points. To the best of our knowledge, we are the first to propose learned scene compression for visual localization. We also demonstrate the effectiveness and efficiency of our method on various outdoor datasets where it can perform accurate localization with low memory consumption.

4	Proceedings of the IEEE CVPR, 2022, pp. 12371-12381	AEGNN: Asynchronous Event-Based Graph Neural Networks	Simon Schaefer, Daniel Gehrig, Davide Scaramuzza	Dept. Informatics, Univ. of Zurich and Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich	相关(GNN)	The best performing learning algorithms devised for event cameras work by first converting events into dense representations that are then processed using standard CNNs. However, these steps discard both the sparsity and high temporal resolution of events, leading to high computational burden and latency. For this reason, recent works have adopted Graph Neural Networks (GNNs), which process events as "static" spatio-temporal graphs, which are inherently "sparse". We take this trend one step further by introducing Asynchronous, Event-based Graph Neural Networks (AEGNNs), a novel event-processing paradigm that generalizes standard GNNs to process events as "evolving" spatio-temporal graphs. AEGNNs follow efficient update rules that restrict recomputation of network activations only to the nodes affected by each new event, thereby significantly reducing both computation and latency for event-by-event processing. AEGNNs are easily trained on synchronous inputs and can be converted to efficient, "asynchronous" networks at test time. We thoroughly validate our method on object classification and detection tasks, where we show an up to a 200-fold reduction in computational complexity (FLOPs), with similar or even better performance than state-of-the-art asynchronous methods. This reduction in computation directly translates to an 8-fold reduction in computational latency when compared to standard GNNs, which opens the door to low-latency event-based processing.
5	Proceedings of the IEEE CVPR, 2022, pp. 6635-6645	ElePose: Unsupervised 3D Human Pose Estimation by Predicting Camera Elevation and Learning Normalizing Flows on 2D Poses	Bastian Wandt, James J. Little, Helge Rhodin	University of British Columbia	不相关	Human pose estimation from single images is a challenging problem that is typically solved by supervised learning. Unfortunately, labeled training data does not yet exist for many human activities since 3D annotation requires dedicated motion capture systems. Therefore, we propose an unsupervised approach that learns to predict a 3D human pose from a single image while only being trained with 2D pose data, which can be crowd-sourced and is already widely available. To this end, we estimate the 3D pose that is most likely over random projections, with the likelihood estimated using normalizing flows on 2D poses. While previous work requires strong priors on camera rotations in the training data set, we learn the distribution of camera angles which significantly improves the performance. Another part of our contribution is to stabilize training with normalizing flows on high-dimensional 3D pose data by first projecting the 2D poses to a linear subspace. We outperform state-of-the-art in unsupervised human pose estimation on the benchmark dataset Human3.6M in all metrics.
6	Proceedings of the IEEE CVPR, 2022, pp. 7803-7812	E-CIR: Event-Enhanced Continuous Intensity Recovery	Chen Song, Qixing Huang, Chandrajit Bajaj	The University of Texas at Austin	不相关	A camera begins to sense light the moment we press the shutter button. During the exposure interval, relative motion between the scene and the camera causes motion blur, a common undesirable visual artifact. This paper presents E-CIR, which converts a blurry image into a sharp video represented as a parametric function from time to intensity. E-CIR leverages events as an auxiliary input. We discuss how to exploit the temporal event structure to construct the parametric bases. We demonstrate how to train a deep learning model to predict the function coefficients. To improve the appearance consistency, we further introduce a refinement module to propagate visual features among consecutive frames. Compared to state-of-the-art event-enhanced deblurring approaches, E-CIR generates smoother and more realistic results. The implementation of E-CIR is available at https://github.com/chensong1995/E-CIR.
7	Proceedings of the IEEE CVPR, 2022, pp. 2416-2425	Connecting the Complementary-View Videos: Joint Camera Identification and Subject Association	Ruize Han, Yiyang Gan, Jiacheng Li, Feifan Wang, Wei Feng, Song Wang	College of Intelligence and Computing, Tianjin University, Tianjin, China	不相关	We attempt to connect the data from complementary views, i.e., top view from drone-mounted cameras in the air, and side view from wearable cameras on the ground. Collaborative analysis of such complementary-view data can facilitate to build the air-ground cooperative visual system for various kinds of applications. This is a very challenging problem due to the large view difference between top and side views. In this paper, we develop a new approach that can simultaneously handle three tasks: i) localizing the side-view camera in the top view; ii) estimating the view direction of the side-view camera; iii) detecting and associating the same subjects on the ground across the complementary views. Our main idea is to explore the spatial position layout of the subjects in two views. In particular, we propose a spatial-aware position representation method to embed the spatial-position distribution of the subjects in different views. We further design a cross-view video collaboration framework composed of a camera identification module and a subject association module to simultaneously perform the above three tasks. We collect a new synthetic dataset consisting of top-view and side-view video sequence pairs for performance evaluation and the experimental results show the effectiveness of the proposed method.
8	Proceedings of the IEEE CVPR, 2022, pp. 3355-3364	Progressive Attention on Multi-Level Dense Difference Maps for Generic Event Boundary Detection	Jiaqi Tang, Zhaoyang Liu, Chen Qian, Wayne Wu, Limin Wang	State Key Laboratory for Novel Software Technology, Nanjing University, China	不相关	Generic event boundary detection is an important yet challenging task in video understanding, which aims at detecting the moments where humans naturally perceive event boundaries. The main challenge of this task is perceiving various temporal variations of diverse event boundaries. To this end, this paper presents an effective and end-to-end learnable framework (DDM-Net). To tackle the diversity and complicated semantics of event boundaries, we make three notable improvements. First, we construct a feature bank to store multi-level features of space and time, prepared for difference calculation at multiple scales. Second, to alleviate inadequate temporal modeling of previous methods, we present dense difference maps (DDM) to comprehensively characterize the motion pattern. Finally, we exploit progressive attention on multi-level DDM to jointly aggregate appearance and motion clues. As a result, DDM-Net respectively achieves a significant boost of 14% and 8% on Kinetics-GEBD and TAPOS benchmark, and outperforms the top-1 winner solution of LOVEU Challenge@CVPR 2021 without bells and whistles. The state-of-the-art result demonstrates the effectiveness of richer motion representa-tion and more sophisticated aggregation, in handling the diversity of generic event boundary detection. The code is made available at https://github.com/MCG-NJU/DDM.
9	Proceedings of the IEEE CVPR, 2022, pp. 17804-17813	TimeReplayer: Unlocking the Potential of Event Cameras for Video Interpolation	Weihua He, Kaichao You, Zhendong Qiao, Xu Jia, Ziyang Zhang, Wenhui Wang, Huchuan Lu, Yaoyuan Wang, Jianxing Liao	Department of Precision Instrument, Tsinghua University	不相关	Recording fast motion in a high FPS (frame-per-second) requires expensive high-speed cameras. As an alternative, interpolating low-FPS videos from commodity cameras has attracted significant attention. If only low-FPS videos are available, motion assumptions (linear or quadratic) are necessary to infer intermediate frames, which fail to model complex motions. Event camera, a new camera with pixels producing events of brightness change at the temporal resolution of \mu s (10^ -6 second), is a game-changing device to enable video interpolation at the presence of arbitrarily complex motion. Since event camera is a novel sensor, its potential has not been fulfilled due to the lack of processing algorithms. The pioneering work Time Lens introduced event cameras to video interpolation by designing optical devices to collect a large amount of paired training data of high-speed frames and events, which is too costly to scale. To fully unlock the potential of event cameras, this paper proposes a novel TimeReplayer algorithm to interpolate videos captured by commodity cameras with events. It is trained in an unsupervised cycle-consistent style, canceling the necessity of high-speed training data and bringing the additional ability of video extrapolation. Its state-of-the-art results and demo videos in supplementary reveal the
10	Proceedings of the IEEE CVPR, 2022, pp. 17844-17853	Optical Flow Estimation for Spiking Camera	Liwen Hu, Rui Zhao, Ziluo Ding, Lei Ma, Boxin Shi, Ruiqin Xiong, Tiejun Huang	NERCVT, School of Computer Science, Peking University	不相关	As a bio-inspired sensor with high temporal resolution, the spiking camera has an enormous potential in real applications, especially for motion estimation in high-speed scenes. However, frame-based and event-based methods are not well suited to spike streams from the spiking camera due to the different data modalities. To this end, we present, SCFlow, a tailored deep learning pipeline to estimate optical flow in high-speed scenes from spike streams. Importantly, a novel input representation is introduced which can adaptively remove the motion blur in spike streams according to the prior motion. Further, for training SCFlow, we synthesize two sets of optical flow data for the spiking camera, SPIkingly Flying Things and Photo-realistic High-speed Motion, denoted as SPIFT and PHM respectively, corresponding to random high-speed and well-designed scenes. Experimental results show that the SCFlow can predict optical flow from spike streams in different high-speed scenes. Moreover, SCFlow shows promising generalization on real spike streams. Codes and datasets refer to https://github.com/Acnext/Optical-Flow-For-Spiking-Camera.

11	Proceedings of the IEEE CVPR, 2022, pp. 12972-12980	Uniform Subdivision of Omnidirectional Camera Space for Efficient Spherical Stereo Matching	Donghun Kang, Hyeonjoong Jang, Jungeon Lee, Chong-Min Kyung, Min H. Kim	KAIST	不相关	Omnidirectional cameras have been used widely to better understand surrounding environments. They are often configured as stereo to estimate depth. However, due to the optics of the fisheye lens, conventional epipolar geometry is inapplicable directly to omnidirectional camera images. Intermediate formats of omnidirectional images, such as equirectangular images, have been used. However, stereo matching performance on these image formats has been lower than the conventional stereo due to severe image distortion near pole regions. In this paper, to address the distortion problem of omnidirection-al images, we devise a novel subdivision scheme of a spherical geodesic grid. This enables more isotropic patch sampling of spherical image information in the omnidirectional camera space. Our spherical geodesic grid is tessellated with an equal-arc subdivision, making the cell sizes and in-between distances as uniform as possible, i.e., the arc length of the spherical grid cell's edges is well regularized. Also, our uniformly tessellated coordinates in a 2D image can be transformed into spherical coordinates via one-to-one mapping, allowing for analytical forward/backward transformation. Our uniform tessellation scheme achieves a higher accuracy of stereo matching than the traditional cylindrical and cubemap-based approaches,
12	Proceedings of the IEEE CVPR, 2022, pp. 8801-8810	Spiking Transformers for Event-Based Single Object Tracking	Jiqing Zhang, Bo Dong, Haiwei Zhang, Jianchuan Ding, Felix Heide, Baocai Yin, Xin Yang	Dalian University of Technology	相关(CNN)	Event-based cameras bring a unique capability to tracking, being able to function in challenging real-world conditions as a direct result of their high temporal resolution and high dynamic range. These imagers capture events asynchronously that encode rich temporal and spatial information. However, effectively extracting this information from events remains an open challenge. In this work, we propose a spiking transformer network, STNet, for single object tracking. STNet dynamically extracts and fuses information from both temporal and spatial domains. In particular, the proposed architecture features a transformer module to provide global spatial information and a spiking neural network (SNN) module for extracting temporal cues. The spiking threshold of the SNN module is dynamically adjusted based on the statistical cues of the spatial information, which we find essential in providing robust SNN features. We fuse both feature branches dynamically with a novel cross-domain attention fusion algorithm. Extensive experiments on three event-based datasets, FE240hz, EED and VisEvent validate that the proposed STNet outperforms existing state-of-the-art methods in both tracking accuracy and speed with a significant margin.
13	Proceedings of the IEEE CVPR, 2022, pp. 20073-20082	UBoCo: Unsupervised Boundary Contrastive Learning for Generic Event Boundary Detection	Hyolim Kang, Jinwoo Kim, Taehyun Kim, Seon Joo Kim	Yonsei University	不相关	Generic Event Boundary Detection (GEBD) is a newly suggested video understanding task that aims to find one level deeper semantic boundaries of events. Bridging the gap between natural human perception and video understanding, it has various potential applications, including interpretable and semantically valid video parsing. Still at an early development stage, existing GEBD solvers are simple extensions of relevant video understanding tasks, disregarding GEBD's distinctive characteristics. In this paper, we propose a novel framework for unsupervised/supervised GEBD, by using the Temporal Self-similarity Matrix (TSM) as the video representation. The new Recursive TSM Parsing (RTP) algorithm exploits local diagonal patterns in TSM to detect boundaries, and it is combined with the Boundary Contrastive (BoCo) loss to train our encoder to generate more informative TSMs. Our framework can be applied to both unsupervised and supervised settings, with both achieving state-of-the-art performance by a huge margin in GEBD benchmark. Especially, our unsupervised method outperforms previous state-of-the-art
14	Proceedings of the IEEE CVPR, 2022, pp. 17745-17754	Ev-TTA: Test-Time Adaptation for Event-Based Object Recognition	Junho Kim, Inwoo Hwang, Young Min Kim	Department of Electrical and Computer Engineering, Seoul National University	不相关	We introduce Ev-TTA, a simple, effective test-time adaptation algorithm for event-based object recognition. While event cameras are proposed to provide measurements of scenes with fast motions or drastic illumination changes, many existing event-based recognition algorithms suffer from performance deterioration under extreme conditions due to significant domain shifts. Ev-TTA mitigates the severe domain gaps by fine-tuning the pre-trained classifiers during the test phase using loss functions inspired by the spatio-temporal characteristics of events. Since the event data is a temporal stream of measurements, our loss function enforces similar predictions for adjacent events to quickly adapt to the changed environment online. Also, we utilize the spatial correlations between two polarities of events to handle noise under extreme illumination, where different polarities of events exhibit distinctive noise distributions. Ev-TTA demonstrates a large amount of performance gain on a wide range of event-based object recognition tasks without extensive additional training. Our formulation can be successfully applied regardless of input representations and further extended into regression tasks. We expect Ev-TTA to provide the key technique to deploy event-based vision algorithms in challenging real-world applications where significant domain shift is inevitable.
15	Proceedings of the IEEE CVPR, 2022, pp. 17263-17272	Topology Preserving Local Road Network Estimation From Single Onboard Camera Image	Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, Luc Van Gool	Computer Vision Lab, ETH Zurich	不相关	Knowledge of the road network topology is crucial for autonomous planning and navigation. Yet, recovering such topology from a single image has only been explored in part. Furthermore, it needs to refer to the ground plane, where also the driving actions are taken. This paper aims at extracting the local road network topology, directly in the bird's-eye-view (BEV), all in a complex urban setting. The only input consists of a single onboard, forward looking camera image. We represent the road topology using a set of directed lane curves and their interactions, which are captured using their intersection points. To better capture topology, we introduce the concept of minimal cycles and their covers. A minimal cycle is the smallest cycle formed by the directed curve segments (between two intersections). The cover is a set of curves whose segments are involved in forming a minimal cycle. We first show that the covers suffice to uniquely represent the road topology. The covers are then used to supervise deep neural networks, along with the lane curve supervision. These learn to predict the road topology from a single input image. The results on the NuScenes and Argoverse benchmarks are significantly better than those obtained with baselines. Code: https://github.com/ybarancan/TopologicalLaneGraph.
16	Proceedings of the IEEE CVPR, 2022, pp. 5791-5801	CamLiFlow: Bidirectional Camera-LiDAR Fusion for Joint Optical Flow and Scene Flow Estimation	Haisong Liu, Tao Lu, Yihui Xu, Jia Liu, Wenjie Li, Lijun Chen	State Key Laboratory for Novel Software Technology, Nanjing University, China	相关(LiDAR/CNN)	In this paper, we study the problem of jointly estimating the optical flow and scene flow from synchronized 2D and 3D data. Previous methods either employ a complex pipeline that splits the joint task into independent stages, or fuse 2D and 3D information in an "early-fusion" or "late-fusion" manner. Such one-size-fits-all approaches suffer from a dilemma of failing to fully utilize the characteristic of each modality or to maximize the inter-modality complementarity. To address the problem, we propose a novel end-to-end framework, called CamLiFlow. It consists of 2D and 3D branches with multiple bidirectional connections between them in specific layers. Different from previous work, we apply a point-based 3D branch to better extract the geometric features and design a symmetric learnable operator to fuse dense image features and sparse point features. Experiments show that CamLiFlow achieves better performance with fewer parameters. Our method ranks 1st on the KITTI Scene Flow benchmark, outperforming the previous art with 1/7 parameters. Code is available at https://github.com/MCG-NJU/CamLiFlow.
17	Proceedings of the IEEE CVPR, 2022, pp. 3407-3417	Learning Motion-Dependent Appearance for High-Fidelity Rendering of Dynamic Humans From a Single Camera	Jae Shin Yoon, Duygu Ceylan, Tuanfeng Y. Wang, Jingwan Lu, Jimei Yang, Zhixin Shu, Hyun Soo Park	University of Minnesota	不相关	Appearance of dressed humans undergoes a complex geometric transformation induced not only by the static pose but also by its dynamics, i.e., there exists a number of cloth geometric configurations given a pose depending on the way it has moved. Such appearance modeling conditioned on motion has been largely neglected in existing human rendering methods, resulting in rendering of physically implausible motion. A key challenge of learning the dynamics of the appearance lies in the requirement of a prohibitively large amount of observations. In this paper, we present a compact motion representation by enforcing equivariance---a representation is expected to be transformed in the way that the pose is transformed. We model an equivariant encoder that can generate the generalizable representation from the spatial and temporal derivatives of the 3D body surface. This learned representation is decoded by a compositional multi-task decoder that renders high fidelity time-varying appearance. Our experiments show that our method can generate a temporally coherent video of dynamic humans for unseen body poses and novel views given

2022	18	Proceedings of the IEEE CVPR, 2022, pp. 17463-17471	Modeling sRGB Camera Noise With Normalizing Flows	Shayan Kousha, Ali Maleky, Michael S. Brown, Marcus A. Brubaker	York University	不相关	Noise modeling and reduction are fundamental tasks in low-level computer vision. They are particularly important for smartphone cameras relying on small sensors that exhibit visually noticeable noise. There has recently been renewed interest in using data-driven approaches to improve camera noise models via neural networks. These data-driven approaches target noise present in the raw-sensor image before it has been processed by the camera's image signal processor (ISP). Modeling noise in the RAW-rgb domain is useful for improving and testing the in-camera denoising algorithm; however, there are situations where the camera's ISP does not apply denoising or additional denoising is desired when the RAW-rgb domain image is no longer available. In such cases, the sensor noise propagates through the ISP to the final rendered image encoded in standard RGB (sRGB). The nonlinear steps on the ISP culminate in a significantly more complex noise distribution in the sRGB domain and existing raw-domain noise models are unable to capture the sRGB noise distribution. We propose a new sRGB-domain noise model based on normalizing flows that is capable of learning the complex noise distribution found in sRGB images under various ISO levels. Our normalizing flows-based approach outperforms other models by a large margin in noise modeling and synthesis tasks. We also show that image denoisers trained on noisy images synthesized with our noise model outperforms those trained with noise from baselines
	19	Proceedings of the IEEE CVPR, 2022, pp. 17824-17833	Reference-Based Video Super-Resolution Using Multi-Camera Video Triplets	Junyong Lee, Myeonghee Lee, Sunghyun Cho, Seungyong Lee	POSTECH	不相关	We propose the first reference-based video super-resolution (RefVSR) approach that utilizes reference videos for high-fidelity results. We focus on RefVSR in a triple-camera setting, where we aim at super-resolving a low-resolution ultra-wide video utilizing wide-angle and telephoto videos. We introduce the first RefVSR network that recurrently aligns and propagates temporal reference features fused with features extracted from low-resolution frames. To facilitate the fusion and propagation of temporal reference features, we propose a propagative temporal fusion module. For learning and evaluation of our network, we present the first RefVSR dataset consisting of triplets of ultra-wide, wide-angle, and telephoto videos concurrently taken from triple cameras of a smartphone. We also propose a two-stage training strategy fully utilizing video triplets in the proposed dataset for real-world 4x video super-resolution. We extensively evaluate our method, and the result shows the state-of-the-art performance in 4x super-resolution.
	20	Proceedings of the IEEE CVPR, 2022, pp. 8866-8875	LMGP: Lifted Multicut Meets Geometry Projections for Multi-Camera Multi-Object Tracking	Duy M. H. Nguyen, Roberto Henschel, Bodo Rosenhahn, Daniel Sonntag, Paul Swoboda	Max Planck Institute for Informatics, Saarland Informatics Campus	不相关	Multi-Camera Multi-Object Tracking is currently drawing attention in the computer vision field due to its superior performance in real-world applications such as video surveillance with crowded scenes or in wide spaces. In this work, we propose a mathematically elegant multi-camera multiple object tracking approach based on a spatial-temporal lifted multicut formulation. Our model utilizes state-of-the-art tracklets produced by single-camera trackers as proposals. As these tracklets may contain ID-Switch errors, we refine them through a novel pre-clustering obtained from 3D geometry projections. As a result, we derive a better tracking graph without ID switches and more precise affinity costs for the data association phase. Tracklets are then matched to multi-camera trajectories by solving a global lifted multicut formulation that incorporates short and long-range temporal interactions on tracklets located in the same camera as well as inter-camera ones. Experimental results on the WildTrack dataset yield near-perfect performance, outperforming state-of-the-art trackers on Campus while being on par on the PETS-09 dataset.
	21	Proceedings of the IEEE CVPR, 2022, pp. 12819-12828	Camera Pose Estimation Using Implicit Distortion Models	Linfei Pan, Marc Pollefeys, Viktor Larsson	ETH Zurich	不相关	Low-dimensional parametric models are the de-facto standard in computer vision for intrinsic camera calibration. These models explicitly describe the mapping between incoming viewing rays and image pixels. In this paper, we explore an alternative approach which implicitly models the lens distortion. The main idea is to replace the parametric model with a regularization term that ensures the latent distortion map varies smoothly throughout the image. The proposed model is effectively parameter-free and allows us to optimize the 6 degree-of-freedom camera pose without explicitly knowing the intrinsic calibration. We show that the method is applicable to a wide selection of cameras with varying distortion and in multiple applications, such as visual localization and structure-from-motion.
	22	Proceedings of the IEEE CVPR, 2022, pp. 15904-15913	High-Fidelity Human Avatars From a Single RGB Camera	Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, Kun Li	Tianjin University, China	不相关	In this paper, we propose a coarse-to-fine framework to reconstruct a personalized high-fidelity human avatar from a monocular video. To deal with the misalignment problem caused by the changed poses and shapes in different frames, we design a dynamic surface network to recover pose-dependent surface deformations, which help to decouple the shape and texture of the person. To cope with the complexity of textures and generate photo-realistic results, we propose a reference-based neural rendering network and exploit a bottom-up sharpening-guided fine-tuning strategy to obtain detailed textures. Our framework also enables photo-realistic novel view/pose synthesis and shape editing applications. Experimental results on both the public dataset and our collected dataset demonstrate that our method outperforms the state-of-the-art methods. The code and dataset will be available at http://cic.tju.edu.cn/faculty/likun/projects/HF-Avatar.
	23	Proceedings of the IEEE CVPR, 2022, pp. 19935-19947	E2(GO)MOTION: Motion Augmented Event Stream for Egocentric Action Recognition	Chiara Plizzari, Mirco Planamente, Gabriele Goletto, Marco Cannici, Emanuele Gusso, Matteo Matteucci, Barbara Caputo	Politecnico di Torino	不相关	Event cameras are novel bio-inspired sensors, which asynchronously capture pixel-level intensity changes in the form of "events". Due to their sensing mechanism, event cameras have little to no motion blur, a very high temporal resolution and require significantly less power and memory than traditional frame-based cameras. These characteristics make them a perfect fit to several real-world applications such as egocentric action recognition on wearable devices, where fast camera motion and limited power challenge traditional vision sensors. However, the ever-growing field of event-based vision has, to date, overlooked the potential of event cameras in such applications. In this paper, we show that event data is a very valuable modality for egocentric action recognition. To do so, we introduce N-EPIC-Kitchens, the first event-based camera extension of the large-scale EPIC-Kitchens dataset. In this context, we propose two strategies: (i) directly processing event-camera data with traditional video-processing architectures (E^2(GO)) and (ii) using event-data to distill optical flow information E^2(GO)MO). On our proposed benchmark, we show that event data provides a comparable performance to RGB and optical flow, yet without any additional flow computation at deploy time, and an improved performance of up to 4% with respect to RGB only information. The N-EPIC-Kitchens dataset is available at https://github.com/EgocentricVision/N-EPIC-Kitchens.
	24	Proceedings of the IEEE CVPR, 2022, pp. 19989-19998	Cross-Modal Background Suppression for Audio-Visual Event Localization	Yan Xia, Zhou Zhao	Zhejiang University	不相关	Audiovisual Event (AVE) localization requires the model to jointly localize an event by observing audio and visual information. However, in unconstrained videos, both information types may be inconsistent or suffer from severe background noise. Hence this paper proposes a novel cross-modal background suppression network for AVE task, operating at the time- and event-level, aiming to improve localiza-tion performance through suppressing asynchronous audiovisual background frames from the examined events and reducing redundant noise. Specifically, the time-level background suppression scheme forces the audio and visual modality to focus on the related information in the temporal dimension that the opposite modality considers essential, and reduces attention to the segments that the other modal considers as background. The event-level background suppression scheme uses the class activation sequences predicted by audio and visual modalities to control the final event category prediction, which can effectively suppress noise events occurring accidentally in a single modality. Furthermore, we introduce a cross-modal gated attention scheme to extract relevant visual regions from complex scenes exploiting both global visual and audio signals. Extensive experiments show our method outperforms the state-of-the-art
	25	Proceedings of the IEEE CVPR, 2022, pp. 20094-20103	Multi-Grained Spatio-Temporal Features Perceived Network for Event-Based Lip-Reading	Ganchao Tan, Yang Wang, Han Han, Yang Cao, Feng Wu, Zheng-Jun Zha	University of Science and Technology of China, Hefei, China	不相关	Automatic lip-reading (ALR) aims to recognize words using visual information from the speaker's lip movements. In this work, we introduce a novel type of sensing device, event cameras, for the task of ALR. Event cameras have both technical and application advantages over conventional cameras for the ALR task because they have higher temporal resolution, less redundant visual information, and lower power consumption. To recognize words from the event data, we propose a novel Multi-grained Spatio-Temporal Features Perceived Network (MSTP) to perceive fine-grained spatio-temporal features from microsecond time-resolved event data. Specifically, a multi-branch network architecture is designed, in which different grained spatio-temporal features are learned by operating at different frame rates. The branch operating on the low frame rate can perceive spatial complete but temporal coarse features. While the branch operating on the high frame rate can perceive spatial coarse but temporal refinement features. And a message flow module is devised to integrate the features from different branches, leading to perceiving more discriminative spatio-temporal features. In addition, we present the first event-based lip-reading dataset (DVS-Lip) captured by the event camera. Experimental results demonstrated the superiority of the proposed model compared to the state-of-the-art event-based action recognition models and video-based lip-reading models.

26	Proceedings of the IEEE CVPR, 2022, pp. 13967-13976	End-to-End Compressed Video Representation Learning for Generic Event Boundary Detection	Congcong Li, Xinyao Wang, Longyin Wen, Dexiang Hong, Tiejian Luo, Libo Zhang	1University of Chinese Academy of Sciences, Beijing, China	不相关	Generic event boundary detection aims to localize the generic, taxonomy-free event boundaries that segment videos into chunks. Existing methods typically require video frames to be decoded before feeding into the network, which demands considerable computational power and storage space. To that end, we propose a new end-to-end compressed video representation learning for event boundary detection that leverages the rich information in the compressed domain, i.e., RGB, motion vectors, residuals, and the internal group of pictures (GOP) structure, without fully decoding the video. Specifically, we first use the ConvNets to extract features of the I-frames in the GOPs. After that, a light-weight spatial-channel compressed encoder is designed to compute the feature representations of the P-frames based on the motion vectors, residuals and representations of their dependent I-frames. A temporal contrastive module is proposed to determine the event boundaries of video sequences. To remedy the ambiguities of annotations and speed up the training process, we use the Gaussian kernel to preprocess the ground-truth event boundaries. Extensive experiments conducted on the Kinetics-GEBD dataset demonstrate that the proposed method achieves comparable results to the state-of-the-art methods with 4.5x faster running speed.
27	Proceedings of the IEEE CVPR, 2022, pp. 17552-17561	Learning To Zoom Inside Camera Imaging Pipeline	Chengzhou Tang, Yuqiang Yang, Bing Zeng, Ping Tan, Shuaicheng Liu	Simon Fraser University	不相关	Existing single image super-resolution methods are either designed for synthetic data, or for real data but in the RGB-to-RGB or the RAW-to-RGB domain. This paper proposes to zoom an image from RAW to RAW inside the camera imaging pipeline. The RAW-to-RAW domain closes the gap between the ideal and the real degradation models. It also excludes the image signal processing pipeline, which refocuses the model learning onto the super-resolution. To these ends, we design a method that receives a low-resolution RAW as the input and estimates the desired higher-resolution RAW jointly with the degradation model. In our method, two convolutional neural networks are learned to constrain the high-resolution image and the degradation model in lower-dimensional subspaces. This subspace constraint converts the ill-posed SISR problem to a well-posed one. To demonstrate the superiority of the proposed method and the RAW-to-RAW domain, we conduct evaluations on the RealSR and the SR-RAW datasets. The results show that our method performs superiorly over the state-of-the-arts both qualitatively and quantitatively, and it also generalizes well and enables zero-shot transfer across different sensors.
28	Proceedings of the IEEE CVPR, 2022, pp. 1172-1181	A Voxel Graph CNN for Object Classification With Event Cameras	Yongjian Deng, Hao Chen, Hai Liu, Youfu Li	College of Computer Science, Beijing University of Technology	相关(CNN)	Event cameras attract researchers' attention due to their low power consumption, high dynamic range, and extremely high temporal resolution. Learning models on event-based object classification have recently achieved massive success by accumulating sparse events into dense frames to apply traditional 2D learning methods. Yet, these approaches necessitate heavy-weight models and are with high computational complexity due to the redundant information introduced by the sparse-to-dense conversion, limiting the potential of event cameras on real-life applications. This study aims to address the core problem of balancing accuracy and model complexity for event-based classification models. To this end, we introduce a novel graph representation for event data to exploit their sparsity better and customi-ze a lightweight voxel graph convolutional neural network (EV-VGCNN) for event-based classification. Specifically, (1) using voxel-wise vertices rather than previous point-wise inputs to explicitly exploit regional 2D semantics of event streams while keeping the sparsity; (2) proposing a multi-scale feature relational layer (MFRL) to extract spatial and motion cues from each vertex discriminatively concern-ing its distances to neighbors. Comprehensive experiments show that our model can advance state-of-the-art classification accuracy with extremely low model complexity (merely 0.84M parameters).
29	Proceedings of the IEEE CVPR, 2022, pp. 8676-8686	Discrete Time Convolution for Fast Event-Based Stereo	Kaixuan Zhang, Kaiwei Che, Jianguo Zhang, Jie Cheng, Ziyang Zhang, Qinghai Guo, Luziwei Leng	Department of Computer Science and Engineering, Southern University of Science and Technology, China	不相关	Inspired by biological retina, dynamical vision sensor transmits events of instantaneous changes of pixel intensity, giving it a series of advantages over traditional frame-based camera, such as high dynamical range, high temporal resolution and low power consumption. However, extracting information from highly asynchronous event data is a challenging task. Inspired by continuous dynamics of biological neuron models, we propose a novel encoding method for sparse events - continuous time convolution (CTC) - which learns to model the spatial feature of the data with intrinsic dynamics. Adopting channel-wise parameterization, temporal dynamics of the model is synchronized on the same feature map and diverges across different ones, enabling it to embed data in a variety of temporal scales. Abstracted from CTC, we further develop discrete time convolution (DTC) which accelerates the process with lower computational cost. We apply these methods to event-based multi-view stereo matching where they surpass state-of-the-art methods on benchmark criteria of the MVSEC dataset. Spatially sparse event data often leads to inaccurate estimation of edges and local contours. To address this problem, we propose a dual-path architecture in which the feature map is complemented by underlying edge information from original events extracted with spatially-adaptive denormalization. We demonstrate the superiority of our model in terms of speed (up to 110 FPS), accuracy and robustness, showing a great potential for real-time fast depth estimation. Finally, we perform experiments on the recent DSEC dataset to demonstrate the general usage of our model.
30	Proceedings of the IEEE CVPR, 2022, pp. 16344-16353	Autofocus for Event Cameras	Shijie Lin, Yinqiang Zhang, Lei Yu, Bin Zhou, Xiaowei Luo, Jia Pan	The University of Hong Kong	不相关	Focus control (FC) is crucial for cameras to capture sharp images in challenging real-world scenarios. The autofocus (AF) facilitates the FC by automatically adjusting the focus settings. However, due to the lack of effective AF methods for the recently introduced event cameras, their FC still relies on naive AF like manual focus adjustments, leading to poor adaptation in challenging real-world conditions. In particular, the inherent differences between event and frame data in terms of sensing modality, noise, temporal resolutions, etc., bring many challenges in designing an effective AF method for event cameras. To address these challenges, we develop a novel event-based autofocus framework consisting of an event-specific focus measure called event rate (ER) and a robust search strategy called event-based golden search (EGS). To verify the performance of our method, we have collected an event-based autofocus dataset (EAD) containing well-synchronized frames, events, and focal positions in a wide variety of challenging scenes with severe lighting and motion conditions. The experiments on this dataset and additional real-world scenarios demonstrated the superiority of our method over state-of-the-art approaches in terms of efficiency and accuracy.
31	Proceedings of the IEEE CVPR, 2022, pp. 20238-20248	Camera-Conditioned Stable Feature Generation for Isolated Camera Supervised Person Re-Identification	Chao Wu, Wenhang Ge, Ancong Wu, Xiaobin Chang	School of Artificial Intelligence, Sun Yat-sen University, China	不相关	To learn camera-view invariant features for person Re-IDentification (Re-ID), the cross-camera image pairs of each person play an important role. However, such cross-view training samples could be unavailable under the ISolated Camera Supervised (ISCS) setting, e.g., a surveillance system deployed across distant scenes. To handle this challenging problem, a new pipeline is introduced by synthesizing the cross-camera samples in the feature space for model training. Specifically, the feature encoder and generator are end-to-end optimized under a novel method, Camera-Conditioned Stable Feature Generation (CCSFG). Its joint learning procedure raises concern on the stability of generative model training. Therefore, a new feature generator, Sigma-Regularized Conditional Variational Autoencoder (Sigma-Reg CVAE), is proposed with theoretical and experimental analysis on its robustness. Extensive experiments on two ISCS person Re-ID datasets demonstrate the superiority of our CCSFG to the competitors.
32	Proceedings of the IEEE CVPR, 2022, pp. 11132-11142	Learning To Detect Scene Landmarks for Camera Localization	Tien Do, Ondrej Miksik, Joseph DeGol, Hyun Soo Park, Sudipta N. Sinha	University of Minnesota	相关(CNN)	Modern camera localization methods that use image retrieval, feature matching, and 3D structure-based pose estimation require long-term storage of numerous scene images or a vast amount of image features. This can make them unsuitable for resource constrained VR/AR devices and also raises serious privacy concerns. We present a new learned camera localization technique that eliminates the need to store features or a detailed 3D point cloud. Our key idea is to implicitly encode the appearance of a sparse yet salient set of 3D scene points into a convolutional neural network (CNN) that can detect these scene points in query images whenever they are visible. We refer to these points as scene landmarks. We also show that a CNN can be trained to regress bearing vectors for such landmarks even when they are not within the camera's field-of-view. We demonstrate that the predicted landmarks yield accurate pose estimates and that our method outperforms DSAC*, the state-of-the-art in learned localization. Furthermore, extending HLoc (an accurate method) by combining its correspondences with our predictions, boosts its accuracy

33	Proceedings of the IEEE CVPR, 2022, pp. 17182-17191	DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection	Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, Mingxing Tan	Johns Hopkins University	相关(LiDAR)	Lidars and cameras are critical sensors that provide complementary information for 3D detection in autonomous driving. While prevalent multi-modal methods simply decorate raw lidar point clouds with camera features and feed them directly to existing 3D detection models, our study shows that fusing camera features with deep lidar features instead of raw points, can lead to better performance. However, as those features are often augmented and aggregated, a key challenge in fusion is how to effectively align the transformed features from two modalities. In this paper, we propose two novel techniques: InverseAug that inverses geometric-related augmentations, e.g., rotation, to enable accurate geometric alignment between lidar points and image pixels, and LearnableAlign that leverages cross-attention to dynamically capture the correlations between image and lidar features during fusion. Based on InverseAug and LearnableAlign, we develop a family of generic multi-modal 3D detection models named DeepFusion, which is more accurate than previous methods. For example, DeepFusion improves PointPillars, CenterPoint, and 3D-MAN baselines on Pedestrian detection for 6.7, 8.9, and 6.2 LEVEL_2 APH, respectively. Notably, our models achieve state-of-the-art performance on Waymo Open Dataset, and show strong model robustness against input corruptions and out-of-distribution data. Code will be publicly available at https://github.com/tensorflow/lingvo.
34	Proceedings of the IEEE CVPR, 2022, pp. 17755-17764	Time Lens++: Event-Based Frame Interpolation With Parametric Non-Linear Flow and Multi-Scale Fusion	Stepan Tulyakov, Alfredo Bochicchio, Daniel Gehrig, Stamatios Georgoulis, Yuanyou Li, Davide Scaramuzza	Huawei Technologies, Zurich Research Center	不相关	Recently, video frame interpolation using a combination of frame- and event-based cameras has surpassed traditional image-based methods both in terms of performance and memory efficiency. However, current methods still suffer from (i) brittle image-level fusion of complementary interpolation results, that fails in the presence of artifacts in the fused image, (ii) potentially temporally inconsistent and inefficient motion estimation procedures, that run for every inserted frame and (iii) low contrast regions that do not trigger events, and thus cause events-only motion estimation to generate artifacts. Moreover, previous methods were only tested on datasets consisting of planar and far-away scenes, which do not capture the full complexity of the real world. In this work, we address the above problems by introducing multi-scale feature-level fusion and computing one-shot non-linear inter-frame motion---which can be efficiently sampled for image warping---from events and images. We also collect the first large-scale events and frames dataset consisting of more than 100 challenging scenes with depth variations, captured with a new experimental setup based on a beamsplitter. We show that our method improves the reconstruction quality by up to 0.2 dB in terms of PSNR and by up to 15% in LPIPS score. Code and dataset will be released
35	Proceedings of the IEEE CVPR, 2022, pp. 17794-17803	Neural Global Shutter: Learn To Restore Video From a Rolling Shutter Camera With Global Reset Feature	Zhixiang Wang, Xiang Ji, Jia-Bin Huang, Shin'ichi Satoh, Xiao Zhou, Yinqiang Zheng	The University of Tokyo	不相关	Most computer vision systems assume distortion-free images as inputs. The widely used rolling-shutter (RS) image sensors, however, suffer from geometric distortion when the camera and object undergo motion during capture. Extensive researches have been conducted on correcting RS distortions. However, most of the existing work relies heavily on the prior assumptions of scenes or motions. Besides, the motion estimation steps are either oversimplified or computationally inefficient due to the heavy flow warping, limiting their applicability. In this paper, we investigate using rolling shutter with a global reset feature (RSGR) to restore clean global shutter (GS) videos. This feature enables us to turn the rectification problem into a deblur-like one, getting rid of inaccurate and costly explicit motion estimation. First, we build an optic system that captures paired RSGR/GS videos. Second, we develop a novel algorithm incorporating spatial and temporal designs to correct the spatial-varying RSGR distortion. Third, we demonstrate that existing image-to-image translation algorithms can recover clean GS videos from distorted RSGR inputs, yet our algorithm achieves the best performance with the specific designs. Our rendered results are not only visually appealing but also beneficial to downstream tasks. Compared to the state-of-the-art RS solution, our RSGR solution is superior in both effectiveness and efficiency. Considering it is easy to realize without changing the hardware, we believe our RSGR solution can potentially replace the RS solution in taking distortion-free videos with low noise and low budget.
36	Proceedings of the IEEE CVPR, 2022, pp. 3594-3604	Event-Based Video Reconstruction via Potential-Assisted Spiking Neural Network	Lin Zhu, Xiao Wang, Yi Chang, Jianing Li, Tiejun Huang, Yonghong Tian	Peking University	相关(SNN)	Neuromorphic vision sensor is a new bio-inspired imaging paradigm that reports asynchronous, continuously per-pixel brightness changes called 'events' with high temporal resolution and high dynamic range. So far, the event-based image reconstruction methods are based on artificial neural networks (ANN) or hand-crafted spatiotemporal smoothing techniques. In this paper, we first implement the image reconstruction work via deep spiking neural network (SNN) architecture. As the bio-inspired neural networks, SNNs operating with asynchronous binary spikes distributed over time, can potentially lead to greater computational efficiency on event-driven hardware. We propose a novel Event-based Video reconstruction framework based on a fully Spiking Neural Network (EVSNN), which utilizes Leaky-Integrate-and-Fire (LIF) neuron and Membrane Potential (MP) neuron. We find that the spiking neurons have the potential to store useful temporal information (memory) to complete such time-dependent tasks. Furthermore, to better utilize the temporal information, we propose a hybrid potential-assisted framework (PA-EVSNN) using the membrane potential of spiking neuron. The proposed neuron is referred as Adaptive Membrane Potential (AMP) neuron, which adaptively updates the membrane potential according to the input spikes. The experimental results demonstrate that our models achieve comparable performance to ANN-based models on IJRR, MVSEC, and HQF datasets. The energy consumptions of EVSNN and PA-EVSNN are 19.36 times and 7.75 times more computationally efficient than their ANN architectures, respectively. The code and pretrained model are available at
37	Proceedings of the IEEE CVPR, 2022, pp. 1090-1099	TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection With Transformers	Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, Chiew-Lan Tai	Hong Kong University of Science and Technology	相关(LiDAR)	LiDAR and camera are two important sensors for 3D object detection in autonomous driving. Despite the increasing popularity of sensor fusion in this field, the robustness against inferior image conditions, e.g., bad illumination and sensor misalignment, is under-explored. Existing fusion methods are easily affected by such conditions, mainly due to a hard association of LiDAR points and image pixels, established by calibration matrices. We propose TransFusion, a robust solution to LiDAR-camera fusion with a soft-association mechanism to handle inferior image conditions. Specifically, our TransFusion consists of convolutional backbones and a detection head based on a transformer decoder. The first layer of the decoder predicts initial bounding boxes from a LiDAR point cloud using a sparse set of object queries, and its second decoder layer adaptively fuses the object queries with useful image features, leveraging both spatial and contextual relationships. The attention mechanism of the transformer enables our model to adaptively determine where and what information should be taken from the image, leading to a robust and effective fusion strategy. We additionally design an image-guided query initialization strategy to deal with objects that are difficult to detect in point clouds. TransFusion achieves state-of-the-art performance on large-scale datasets. We provide extensive experiments to demonstrate its robustness against degenerated image quality and calibration errors. We also extend the proposed method to the 3D tracking task and achieve the 1st place in the leaderboard of nuScenes tracking, showing its effectiveness and generalization capability.
38	Proceedings of the IEEE CVPR, 2022, pp. 5781-5790	Event-Aided Direct Sparse Odometry	Javier Hidalgo-Carrió, Guillermo Gallego, Davide Scaramuzza	Dept. of Informatics, Univ. of Zurich and Dept. of Neuroinformatics, Univ. of Zurich and ETH Zurich	不相关	We introduce EDS, a direct monocular visual odometry using events and frames. Our algorithm leverages the event generation model to track the camera motion in the blind time between frames. The method formulates a direct probabilistic approach of observed brightness increments. Per-pixel brightness increments are predicted using a sparse number of selected 3D points and are compared to the events via the brightness increment error to estimate camera motion. The method recovers a semi-dense 3D map using photometric bundle adjustment. EDS is the first method to perform 6-DOF VO using events and frames with a direct approach. By design it overcomes the problem of changing appearance in indirect methods. Our results outperform all previous event-based odometry solutions. We also show that, for a target error performance, EDS can work at lower frame rates than state-of-the-art frame-based VO solutions. This opens the door to low-power motion-tracking applications where frames are sparingly triggered "on demand" and our method tracks the motion in between. We release code and datasets to the public.

	39	Proceedings of the IEEE CVPR, 2022, pp. 16420-16429	CLIP-Event: Connecting Text and Images With Event Structures	Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, Shih-Fu Chang	University of Illinois Urbana-Champaign	不相关	Vision-language (V+L) pretraining models have achieved great success in supporting multimedia applications by understanding the alignments between images and text. While existing vision-language pretraining models primarily focus on understanding objects in images or entities in text, they often ignore the alignment at the level of events and their argument structures. In this work, we propose a contrastive learning framework to enforce vision-language pretraining models to comprehend events and associated argument (participant) roles. To achieve this, we take advantage of text information extraction technologies to obtain event structural knowledge, and utilize multiple prompt functions to contrast difficult negative descriptions by manipulating event structures. We also design an event graph alignment loss based on optimal transport to capture event argument structures. In addition, we collect a large event-rich dataset (106,875 images) for pretraining, which provides a more challenging image retrieval benchmark to assess the understanding of complicated lengthy sentences. Experiments show that our zero-shot CLIP-Event outperforms the state-of-the-art supervised model in argument extraction on Multimedia Event Extraction, achieving more than 5% absolute F-score gain in event extraction, as well as significant improvements on a variety of downstream tasks under zero-shot settings.
	1	Proceedings of the IEEE CVPR, 2023, pp.14592-14601	Video Event Restoration Based on Keyframes for Video Anomaly Detection	Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, Xiaotao Liu	Guangzhou Institute of Technology, Xidian University, Guangzhou, China	相关(DNN)	Video anomaly detection (VAD) is a significant computer vision problem. Existing deep neural network (DNN) based VAD methods mostly follow the route of frame reconstruction or frame prediction. However, the lack of mining and learning of higher-level visual features and temporal context relationships in videos limits the further performance of these two approaches. Inspired by video codec theory, we introduce a brand-new VAD paradigm to break through these limitations: First, we propose a new task of video event restoration based on keyframes. Encouraging DNN to infer missing multiple frames based on video keyframes so as to restore a video event, which can more effectively motivate DNN to mine and learn potential higher-level visual features and comprehensive temporal context relationships in the video. To this end, we propose a novel U-shaped Swin Transformer Network with Dual Skip Connections (USTN-DSC) for video event restoration, where a cross-attention and a temporal upsampling residual skip connection are introduced to further assist in restoring complex static and dynamic motion object features in the video. In addition, we propose a simple and effective adjacent frame difference loss to constrain the motion consistency of the video sequence. Extensive experiments on benchmarks demonstrate that USTN-DSC outperforms most existing methods, validating the effectiveness of our method.
	2	Proceedings of the IEEE CVPR, 2023, pp. 9771-9780	Adaptive Global Decay Process for Event Cameras	Urbano Miguel Nunes, Ryad Benosman, Sio-Hoi Ieng	Sorbonne University, 4 place jussieu, 75005 Paris	不相关	In virtually all event-based vision problems, there is the need to select the most recent events, which are assumed to carry the most relevant information content. To achieve this, at least one of three main strategies is applied, namely: 1) constant temporal decay or fixed time window, 2) constant number of events, and 3) flow-based lifetime of events. However, these strategies suffer from at least one major limitation each. We instead propose a novel decay process for event cameras that adapts to the global scene dynamics and whose latency is in the order of nanoseconds. The main idea is to construct an adaptive quantity that encodes the global scene dynamics, denoted by event activity. The proposed method is evaluated in several event-based vision problems and datasets, consistently improving the corresponding baseline methods' performance. We thus believe it can have a significant widespread impact on event-based research. Code available:
	3	Proceedings of the IEEE CVPR, 2023, pp.22867-22876	Hierarchical Neural Memory Network for Low Latency Event Processing	Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, Ken Sakurada	National Institute of Advanced Industrial Science and Technology (AIST)	相关(CNN)	This paper proposes a low latency neural network architecture for event-based dense prediction tasks. Conventional architectures encode entire scene contents at a fixed rate regardless of their temporal characteristics. Instead, the proposed network encodes contents at a proper temporal scale depending on its movement speed. We achieve this by constructing temporal hierarchy using stacked latent memories that operate at different rates. Given low latency event steams, the multi-level memories gradually extract dynamic to static scene contents by propagating information from the fast to the slow memory modules. The architecture not only reduces the redundancy of conventional architectures but also exploits long-term dependencies. Furthermore, an attention-based event representation efficiently encodes sparse event streams into the memory cells. We conduct extensive evaluations on three event-based dense prediction tasks, where the proposed approach outperforms the existing methods on accuracy and latency, while demonstrating effective event and image fusion capabilities. The code is available at https://hamarh.github.io/hmnet/
	4	Proceedings of the IEEE CVPR, 2023, pp.21940-21949	Tangentially Elongated Gaussian Belief Propagation for Event-Based Incremental Optical Flow Estimation	Jun Nagata, Yusuke Sekikawa	DENSO IT LAB., INC., Japan.	相关(DNN)	Optical flow estimation is a fundamental functionality in computer vision. An event-based camera, which asynchronously detects sparse intensity changes, is an ideal device for realizing low-latency estimation of the optical flow owing to its low-latency sensing mechanism. An existing method using local plane fitting of events could utilize the sparsity to realize incremental updates for low-latency estimation; however, its output is merely a normal component of the full optical flow. An alternative approach using a frame-based deep neural network could estimate the full flow; however, its intensive non-incremental dense operation prohibits the low-latency estimation. We propose tangentially elongated Gaussian (TEG) belief propagation (BP) that realizes incremental full-flow estimation. We model the probability of full flow as the joint distribution of TEGs from the normal flow measurements, such that the marginal of this distribution with correct prior equals the full flow. We formulate the marginalization using a message-passing based on the BP to realize efficient incremental updates using sparse measurements. In addition to the theoretical justification, we evaluate the effectiveness of the TEGBP in real-world datasets; it outperforms SOTA incremental quasi-full flow method by a large margin. The code will be open-
	5	Proceedings of the IEEE CVPR, 2023, pp.17797-17807	Learning Adaptive Dense Event Stereo From the Image Domain	Hoonhee Cho, Jegyeong Cho, Kuk-Jin Yoon	Visual Intelligence Lab., KAIST, Korea	不相关	Recently, event-based stereo matching has been studied due to its robustness in poor light conditions. However, existing event-based stereo networks suffer severe performance degradation when domains shift. Unsupervised domain adaptation (UDA) aims at resolving this problem without using the target domain ground-truth. However, traditional UDA still needs the input event data with ground-truth in the source domain, which is more challenging and costly to obtain than image data. To tackle this issue, we propose a novel unsupervised domain Adaptive Dense Event Stereo (ADES), which resolves gaps between the different domains and input modalities. The proposed ADES framework adapts event-based stereo networks from abundant image datasets with ground-truth on the source domain to event datasets without ground-truth on the target domain, which is a more practical setup. First, we propose a self-supervision module that trains the network on the target domain through image reconstruction, while an artifact prediction network trained on the source domain assists in removing intermittent artifacts in the reconstructed image. Secondly, we utilize the feature-level normalization scheme to align the extracted features along the epipolar line. Finally, we present the motion-invariant consistency module to impose the consistent output between the perturbed motion. Our experiments demonstrate that our approach achieves remarkable results in the adaptation ability of event-based stereo matching from the
	6	Proceedings of the IEEE CVPR, 2023, pp.4150-4159	F2-NeRF: Fast Neural Radiance Field Training With Free Camera Trajectories	Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, Wenping Wang	The University of Hong Kong	不相关	This paper presents a novel grid-based NeRF called F^2-NeRF (Fast-Free-NeRF) for novel view synthesis, which enables arbitrary input camera trajectories and only costs a few minutes for training. Existing fast grid-based NeRF training frameworks, like Instant-NGP, Plenoxels, DVGO, or TensoRF, are mainly designed for bounded scenes and rely on space warping to handle unbounded scenes. Existing two widely-used space-warping methods are only designed for the forward-facing trajectory or the 360deg object-centric trajectory but cannot process arbitrary trajectories. In this paper, we delve deep into the mechanism of space warping to handle unbounded scenes. Based on our analysis, we further propose a novel space-warping method called perspective warping, which allows us to handle arbitrary trajectories in the grid-based NeRF framework. Extensive experiments demonstrate that F^2-NeRF is able to use the same perspective warping to render high-quality images on two standard datasets and a new free trajectory dataset collected by us.

7	Proceedings of the IEEE CVPR, 2023, pp.18827-18836	Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception	Junyu Gao, Mengyuan Chen, Changsheng Xu	State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS)	不相关	With only video-level event labels, this paper targets at the task of weakly-supervised audio-visual event perception (WS-AVEP), which aims to temporally localize and categorize events belonging to each modality. Despite the recent progress, most existing approaches either ignore the unsynchronized property of audio-visual tracks or discount the complementary modality for explicit enhancement. We argue that, for an event residing in one modality, the modality itself should provide ample presence evidence of this event, while the other complementary modality is encouraged to afford the absence evidence as a reference signal. To this end, we propose to collect Cross-Modal Presence-Absence Evidence (CMPAE) in a unified framework. Specifically, by leveraging uni-modal and cross-modal representations, a presence-absence evidence collector (PAEC) is designed under Subjective Logic theory. To learn the evidence in a reliable range, we propose a joint-modal mutual learning (JML) process, which calibrates the evidence of diverse audible, visible, and audi-visible events adaptively and dynamically. Extensive experiments show that our method surpasses state-of-the-arts (e.g., absolute gains of 3.6% and 6.1% in terms of event-level visual and audio metrics). Code is available in github.com/MengyuanChen21/CVPR2023-
8	Proceedings of the IEEE CVPR, 2023, pp.17307-17316	Perspective Fields for Single Image Camera Calibration	Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, David F. Fouhey	University of Michigan	相关(CNN)	Geometric camera calibration is often required for applications that understand the perspective of the image. We propose perspective fields as a representation that models the local perspective properties of an image. Perspective Fields contain per-pixel information about the camera view, parameterized as an up vector and a latitude value. This representation has a number of advantages as it makes minimal assumptions about the camera model and is invariant or equivariant to common image editing operations like cropping, warping, and rotation. It is also more interpretable and aligned with human perception. We train a neural network to predict Perspective Fields and the predicted Perspective Fields can be converted to calibration parameters easily. We demonstrate the robustness of our approach under various scenarios compared with camera calibration-based methods and show example applications in image compositing. Project page:
9	Proceedings of the IEEE CVPR, 2023, pp.9243-9252	Collaboration Helps Camera Overtake LiDAR in 3D Detection	Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, Yanfeng Wang	Cooperative Medianet Innovation Center, Shanghai Jiao Tong University	相关(LiDAR/NN)	Camera-only 3D detection provides an economical solution with a simple configuration for localizing objects in 3D space compared to LiDAR-based detection systems. However, a major challenge lies in precise depth estimation due to the lack of direct 3D measurements in the input. Many previous methods attempt to improve depth estimation through network designs, e.g., deformable layers and larger receptive fields. This work proposes an orthogonal direction, improving the camera-only 3D detection by introducing multi-agent collaborations. Our proposed collaborative camera-only 3D detection (CoCa3D) enables agents to share complementary information with each other through communication. Meanwhile, we optimize communication efficiency by selecting the most informative cues. The shared messages from multiple viewpoints disambiguate the single-agent estimated depth and complement the occluded and long-range regions in the single-agent view. We evaluate CoCa3D in one real-world dataset and two new simulation datasets. Results show that CoCa3D improves previous SOTA performances by 44.21% on DAIR-V2X, 30.60% on OPV2V+, 12.59% on CoPerception-UAVs+ for AP@70. Our preliminary results show a potential that with sufficient collaboration, the camera might overtake LiDAR in some practical scenarios. We released the dataset and code at https://siheng-chen.github.io/dataset/CoPerception+ and https://github.com/MediaBrain-SJTU/CoCa3D.
10	Proceedings of the IEEE CVPR, 2023, pp.21307-21316	SMOC-Net: Leveraging Camera Pose for Self-Supervised Monocular Object Pose Estimation	Tao Tan, Qiulei Dong	School of Artificial Intelligence, UCAS	相关(CNN)	Recently, self-supervised 6D object pose estimation, where synthetic images with object poses (sometimes jointly with un-annotated real images) are used for training, has attracted much attention in computer vision. Some typical works in literature employ a time-consuming differentiable renderer for object pose prediction at the training stage, so that (i) their performances on real images are generally limited due to the gap between their rendered images and real images and (ii) their training process is computationally expensive. To address the two problems, we propose a novel Network for Self-supervised Monocular Object pose estimation by utilizing the predicted Camera poses from un-annotated real images, called SMOC-Net. The proposed network is explored under a knowledge distillation framework, consisting of a teacher model and a student model. The teacher model contains a backbone estimation module for initial object pose estimation, and an object pose refiner for refining the initial object poses using a geometric constraint (called relative-pose constraint) derived from relative camera poses. The student model gains knowledge for object pose estimation from the teacher model by imposing the relative-pose constraint. Thanks to the relative-pose constraint, SMOC-Net could not only narrow the domain gap between synthetic and real data but also reduce the training cost. Experimental results on two public datasets demonstrate that SMOC-Net outperforms several state-of-the-art methods by a large margin while requiring much less training time than the differentiable-renderer-based methods.
11	Proceedings of the IEEE CVPR, 2023, pp.1002-1011	BEV-LaneDet: An Efficient 3D Lane Detection Based on Virtual Camera via Key-Points	Ruihao Wang, Jian Qin, Kaiying Li, Yaochen Li, Dong Cao, Jintao Xu	HAOMO.AI Technology Co., Ltd	相关(DNN)	3D lane detection which plays a crucial role in vehicle routing, has recently been a rapidly developing topic in autonomous driving. Previous works struggle with practicality due to their complicated spatial transformations and inflexible representations of 3D lanes. Faced with the issues, our work proposes an efficient and robust monocular 3D lane detection called BEV-LaneDet with three main contributions. First, we introduce the Virtual Camera that unifies the in/extrinsic parameters of cameras mounted on different vehicles to guarantee the consistency of the spatial relationship among cameras. It can effectively promote the learning procedure due to the unified visual space. We secondly propose a simple but efficient 3D lane representation called Key-Points Representation. This module is more suitable to represent the complicated and diverse 3D lane structures. At last, we present a light-weight and chip-friendly spatial transformation module named Spatial Transformation Pyramid to transform multiscale front-view features into BEV features. Experimental results demonstrate that our work outperforms the state-of-the-art approaches in terms of F-Score, being 10.6% higher on the OpenLane dataset and 4.0% higher on the Apollo 3D synthetic dataset, with a speed of 185 FPS. Code is released at https://github.com/gigo-
12	Proceedings of the IEEE CVPR, 2023, pp.9781-9790	Frame-Event Alignment and Fusion Network for High Frame Rate Tracking	Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, Xin Yang	Dalian University of Technology	不相关	Most existing RGB-based trackers target low frame rate benchmarks of around 30 frames per second. This setting restricts the tracker's functionality in the real world, especially for fast motion. Event-based cameras as bioinspired sensors provide considerable potential for high frame rate tracking due to their high temporal resolution. However, event-based cameras cannot offer fine-grained texture information like conventional cameras. This unique complementarity motivates us to combine conventional frames and events for high frame rate object tracking under various challenging conditions. In this paper, we propose an end-to-end network consisting of multi-modality alignment and fusion modules to effectively combine meaningful information from both modalities at different measurement rates. The alignment module is responsible for cross-modality and cross-frame-rate alignment between frame and event modalities under the guidance of the moving cues furnished by events. While the fusion module is accountable for emphasizing valuable features and suppressing noise information by the mutual complement between the two modalities. Extensive experiments show that the proposed approach outperforms state-of-the-art trackers by a significant margin in high frame rate tracking. With the FE240hz dataset, our
13	Proceedings of the IEEE CVPR, 2023, pp.18032-18042	Event-Based Video Frame Interpolation With Cross-Modal Asymmetric Bidirectional Motion Fields	Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, Kuk-Jin Yoon	Korea Advanced Institute of Science and Technology	不相关	Video Frame Interpolation (VFI) aims to generate intermediate video frames between consecutive input frames. Since the event cameras are bio-inspired sensors that only encode brightness changes with a micro-second temporal resolution, several works utilized the event camera to enhance the performance of VFI. However, existing methods estimate bidirectional inter-frame motion fields with only events or approximations, which can not consider the complex motion in real-world scenarios. In this paper, we propose a novel event-based VFI framework with cross-modal asymmetric bidirectional motion field estimation. In detail, our EIF-BiOFNet utilizes each valuable characteristic of the events and images for direct estimation of inter-frame motion fields without any approximation methods.Moreover, we develop an interactive attention-based frame synthesis network to efficiently leverage the complementary warping-based and synthesis-based features. Finally, we build a large-scale event-based VFI dataset, ERF-X170FPS, with a high frame rate, extreme motion, and dynamic textures to overcome the limitations of previous event-based VFI datasets. Extensive experimental results validate that our method shows significant performance improvement over the state-of-the-art VFI methods on various datasets.Our project
14	Proceedings of the IEEE CVPR, 2023, pp.20616-20625	High-Fidelity Event-Radiance Recovery via Transient Event Frequency	Jin Han, Yuta Asano, Boxin Shi, Yinqiang Zheng, Imari Sato	Graduate School of Information Science and Technology, The University of Tokyo	不相关	High-fidelity radiance recovery plays a crucial role in scene information reconstruction and understanding. Conventional cameras suffer from limited sensitivity in dynamic range, bit depth, and spectral response, etc. In this paper, we propose to use event cameras with bio-inspired silicon sensors, which are sensitive to radiance changes, to recover precise radiance values. We reveal that, under active lighting conditions, the transient frequency of event signals triggering linearly reflects the radiance value. We propose an innovative method to convert the high temporal resolution of event signals into precise radiance values. The precise radiance values yields several capabilities in image analysis. We demonstrate the feasibility of recovering radiance values solely from the transient event frequency (TEF) through multiple experiments.

15	Proceedings of the IEEE CVPR, 2023, pp.8917-8926	Robot Structure Prior Guided Temporal Attention for Camera-to-Robot Pose Estimation From Image Sequence	Yang Tian, Jiyao Zhang, Zekai Yin, Hao Dong	CFCS, Peking University	不相关	In this work, we tackle the problem of online camera-to-robot pose estimation from single-view successive frames of an image sequence, a crucial task for robots to interact with the world. The primary obstacles of this task are the robot's self-occlusions and the ambiguity of single-view images. This work demonstrates, for the first time, the effectiveness of temporal information and the robot structure prior in addressing these challenges. Given the successive frames and the robot joint configuration, our method learns to accurately regress the 2D coordinates of the predefined robot's keypoints (e.g., joints). With the camera intrinsic and robotic joints status known, we get the camera-to-robot pose using a Perspective-n-point (PnP) solver. We further improve the camera-to-robot pose iteratively using the robot structure prior. To train the whole pipeline, we build a large-scale synthetic dataset generated with domain randomisation to bridge the sim-to-real gap. The extensive experiments on synthetic and real-world datasets and the downstream robotic grasping task demonstrate that our method achieves new state-of-the-art performances and outperforms traditional hand-eye calibration algorithms in real-time (36 FPS). Code and data are available at the project page: https://sites.google.com/view/sgtapose.
16	Proceedings of the IEEE CVPR, 2023, pp.4992-5002	EventNeRF: Neural Radiance Fields From a Single Colour Event Camera	Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, Vladislav Golyanik	Max Planck Institute for Informatics, SIC	不相关	Asynchronously operating event cameras find many applications due to their high dynamic range, vanishingly low motion blur, low latency and low data bandwidth. The field saw remarkable progress during the last few years, and existing event-based 3D reconstruction approaches recover sparse point clouds of the scene. However, such sparsity is a limiting factor in many cases, especially in computer vision and graphics, that has not been addressed satisfactorily so far. Accordingly, this paper proposes the first approach for 3D-consistent, dense and photorealistic novel view synthesis using just a single colour event stream as input. At its core is a neural radiance field trained entirely in a self-supervised manner from events while preserving the original resolution of the colour event channels. Next, our ray sampling strategy is tailored to events and allows for data-efficient training. At test, our method produces results in the RGB space at unprecedented quality. We evaluate our method qualitatively and numerically on several challenging synthetic and real scenes and show that it produces significantly denser and more visually appealing renderings than the existing methods. We also demonstrate robustness in challenging scenarios with fast motion and under low lighting conditions. We release the newly recorded dataset and our source code to facilitate the research field, see https://4dqv.mpi-inf.mpg.de/EventNeRF.
17	Proceedings of the IEEE CVPR, 2023, pp.1537-1546	Progressive Spatio-Temporal Alignment for Efficient Event-Based Motion Estimation	Xueyan Huang, Yueyi Zhang, Zhiwei Xiong	University of Science and Technology of China	不相关	In this paper, we propose an efficient event-based motion estimation framework for various motion models. Different from previous works, we design a progressive event-to-map alignment scheme and utilize the spatio-temporal correlations to align events. In detail, we progressively align sampled events in an event batch to the time-surface map and obtain the updated motion model by minimizing a novel time-surface loss. In addition, a dynamic batch size strategy is applied to adaptively adjust the batch size so that all events in the batch are consistent with the current motion model. Our framework has three advantages: a) the progressive scheme refines motion parameters iteratively, achieving accurate motion estimation; b) within one iteration, only a small portion of events are involved in optimization, which greatly reduces the total runtime; c) the dynamic batch size strategy ensures that the constant velocity assumption always holds. We conduct comprehensive experiments to evaluate our framework on challenging high-speed scenes with three motion models: rotational, homography, and 6-DOF models. Experimental results demonstrate that our framework achieves state-of-the-art estimation accuracy and efficiency.
18	Proceedings of the IEEE CVPR, 2023, pp.1547-1556	Event-Based Shape From Polarization	Manasi Muglikar, Leonard Bauersfeld, Diederik Paul Moeys, Davide Scaramuzza	Robotics and Perception Group, University of Zurich, Switzerland	不相关	State-of-the-art solutions for Shape-from-Polarization (SfP) suffer from a speed-resolution tradeoff: they either sacrifice the number of polarization angles measured or necessitate lengthy acquisition times due to framerate constraints, thus compromising either accuracy or latency. We tackle this tradeoff using event cameras. Event cameras operate at microseconds resolution with negligible motion blur, and output a continuous stream of events that precisely measures how light changes over time asynchronously. We propose a setup that consists of a linear polarizer rotating at high speeds in front of an event camera. Our method uses the continuous event stream caused by the rotation to reconstruct relative intensities at multiple polarizer angles. Experiments demonstrate that our method outperforms physics-based baselines using frames, reducing the MAE by 25% in synthetic and real-world datasets. In the real world, we observe, however, that the challenging conditions (i.e., when few events are generated) harm the performance of physics-based solutions. To overcome this, we propose a learning-based approach that learns to estimate surface normals even at low event-rates, improving the physics-based approach by 52% on the real world dataset. The proposed system achieves an acquisition speed equivalent to 50 fps (>twice the framerate of the commercial polarization sensor) while retaining the spatial resolution of 1MP. Our evaluation is based on the first large-scale dataset for event-based SfP.
19	Proceedings of the IEEE CVPR, 2023, pp.6945-6956	EXIF As Language: Learning Cross-Modal Associations Between Images and Camera Metadata	Chenhao Zheng, Ayush Shrivastava, Andrew Owens	University of Michigan	不相关	We learn a visual representation that captures information about the camera that recorded a given photo. To do this, we train a multimodal embedding between image patches and the EXIF metadata that cameras automatically insert into image files. Our model represents this metadata by simply converting it to text and then processing it with a transformer. The features that we learn significantly outperform other self-supervised and supervised features on downstream image forensics and calibration tasks. In particular, we successfully localize spliced image regions "zero shot" by clustering the visual embeddings for all of the patches within an image.
20	Proceedings of the IEEE CVPR, 2023, pp.17928-17938	Standing Between Past and Future: Spatio-Temporal Modeling for Multi-Camera 3D Multi-Object Tracking	Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, Yu-Xiong Wang	University of Illinois Urbana-Champaign	不相关	This work proposes an end-to-end multi-camera 3D multi-object tracking (MOT) framework. It emphasizes spatio-temporal continuity and integrates both past and future reasoning for tracked objects. Thus, we name it "Past-and-Future reasoning for Tracking" (PF-Track). Specifically, our method adapts the "tracking by attention" framework and represents tracked instances coherently over time with object queries. To explicitly use historical cues, our "Past Reasoning" module learns to refine the tracks and enhance the object features by cross-attending to queries from previous frames and other objects. The "Future Reasoning" module digests historical information and predicts robust future trajectories. In the case of long-term occlusions, our method maintains the object positions and enables re-association by integrating motion predictions. On the nuScenes dataset, our method improves AMOTA by a large margin and remarkably reduces ID-Switches by 90% compared to prior approaches, which is an order of magnitude less. The code and models are made available at https://github.com/TRI-ML/PF-Track.
21	Proceedings of the IEEE CVPR, 2023, pp.21570-21579	CAPE: Camera View Position Embedding for Multi-View 3D Object Detection	Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang, Xiang Bai	Huazhong University of Science and Technology	不相关	In this paper, we address the problem of detecting 3D objects from multi-view images. Current query-based methods rely on global 3D position embeddings (PE) to learn the geometric correspondence between images and 3D space. We claim that directly interacting 2D image features with global 3D PE could increase the difficulty of learning view transformation due to the variation of camera extrinsics. Thus we propose a novel method based on CAmera view Position Embedding, called CAPE. We form the 3D position embeddings under the local camera-view coordinate system instead of the global coordinate system, such that 3D position embedding is free of encoding camera extrinsic parameters. Furthermore, we extend our CAPE to temporal modeling by exploiting the object queries of previous frames and encoding the ego motion for boosting 3D object detection. CAPE achieves the state-of-the-art performance (61.0% NDS and 52.5% mAP) among all LiDAR-free methods on standard nuScenes dataset. Codes and models are available.
22	Proceedings of the IEEE CVPR, 2023, pp.13884-13893	Recurrent Vision Transformers for Object Detection With Event Cameras	Mathias Gehrig, Davide Scaramuzza	Robotics and Perception Group, University of Zurich, Switzerland	相关(DNN/SNN)	We present Recurrent Vision Transformers (RVTs), a novel backbone for object detection with event cameras. Event cameras provide visual information with sub-millisecond latency at a high-dynamic range and with strong robustness against motion blur. These unique properties offer great potential for low-latency object detection and tracking in time-critical scenarios. Prior work in event-based vision has achieved outstanding detection performance but at the cost of substantial inference time, typically beyond 40 milliseconds. By revisiting the high-level design of recurrent vision backbones, we reduce inference time by a factor of 6 while retaining similar performance. To achieve this, we explore a multi-stage design that utilizes three key concepts in each stage: First, a convolutional prior that can be regarded as a conditional positional embedding. Second, local- and dilated global self-attention for spatial feature interaction. Third, recurrent temporal feature aggregation to minimize latency while retaining temporal information. RVTs can be trained from scratch to reach state-of-the-art performance on event-based object detection - achieving an mAP of 47.2% on the Gen1 automotive dataset. At the same time, RVTs offer fast inference (<12 ms on a T4 GPU) and favorable parameter efficiency (5 times fewer than prior art). Our study brings new insights into effective design choices that can be fruitful for research beyond event-based vision.

2023	23	Proceedings of the IEEE CVPR, 2023, pp.21488-21497	DC2: Dual-Camera Defocus Control by Learning To Refocus	Hadi Alzayer, Abdullah Abuolaim, Leung Chun Chan, Yang Yang, Ying Chen Lou, Jia-Bin Huang, Abhishek Kar	Google	不相关	Smartphone cameras today are increasingly approaching the versatility and quality of professional cameras through a combination of hardware and software advancements. However, fixed aperture remains a key limitation, preventing users from controlling the depth of field (DoF) of captured images. At the same time, many smartphones now have multiple cameras with different fixed apertures - specifically, an ultra-wide camera with wider field of view and deeper DoF and a higher resolution primary camera with shallower DoF. In this work, we propose DC^2, a system for defocus control for synthetically varying camera aperture, focus distance and arbitrary defocus effects by fusing information from such a dual-camera system. Our key insight is to leverage real-world smartphone camera dataset by using image refocus as a proxy task for learning to control defocus. Quantitative and qualitative evaluations on real-world data demonstrate our system's efficacy where we outperform state-of-the-art on defocus deblurring, bokeh rendering, and image refocus. Finally, we demonstrate creative post-capture defocus control enabled by our method, including tilt-shift and content-based defocus effects.
	24	Proceedings of the IEEE CVPR, 2023, pp.21222-21232	Decoupling Human and Camera Motion From Videos in the Wild	Vickie Ye, Georgios Pavlakos, Jitendra Malik, Angjoo Kanazawa	University of California, Berkeley	不相关	We propose a method to reconstruct global human trajectories from videos in the wild. Our optimization method decouples the camera and human motion, which allows us to place people in the same world coordinate frame. Most existing methods do not model the camera motion; methods that rely on the background pixels to infer 3D human motion usually require a full scene reconstruction, which is often not possible for in-the-wild videos. However, even when existing SLAM systems cannot recover accurate scene reconstructions, the background pixel motion still provides enough signal to constrain the camera motion. We show that relative camera estimates along with data-driven human motion priors can resolve the scene scale ambiguity and recover global human trajectories. Our method robustly recovers the global 3D trajectories of people in challenging in-the-wild videos, such as PoseTrack. We quantify our improvement over existing methods on 3D human dataset Egobody. We further demonstrate that our recovered camera scale allows us to reason about motion of multiple people in a shared coordinate frame, which improves performance of downstream tracking in PoseTrack. Code and additional results can be found at https://vye16.github.io/slahmr/.
	25	Proceedings of the IEEE CVPR, 2023, pp.5186-5195	LiDAR2Map: In Defense of LiDAR-Based Semantic Map Construction Using Online Camera Distillation	Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, Jianke Zhu	Zhejiang University	相关(LiDAR)	Semantic map construction under bird's-eye view (BEV) plays an essential role in autonomous driving. In contrast to camera image, LiDAR provides the accurate 3D observations to project the captured 3D features onto BEV space inherently. However, the vanilla LiDAR-based BEV feature often contains many indefinite noises, where the spatial features have little texture and semantic cues. In this paper, we propose an effective LiDAR-based method to build semantic map. Specifically, we introduce a BEV pyramid feature decoder that learns the robust multi-scale BEV features for semantic map construction, which greatly boosts the accuracy of the LiDAR-based method. To mitigate the defects caused by lacking semantic cues in LiDAR data, we present an online Camera-to-LiDAR distillation scheme to facilitate the semantic learning from image to point cloud. Our distillation scheme consists of feature-level and logit-level distillation to absorb the semantic information from camera in BEV. The experimental results on challenging nuScenes dataset demonstrate the efficacy of our proposed LiDAR2Map on semantic map construction, which significantly outperforms the previous LiDAR-based methods over 27.9% mIoU and even performs better than the state-of-the-art camera-based approaches. Source code is available at:
	26	Proceedings of the IEEE CVPR, 2023, pp.21349-21359	SparsePose: Sparse-View Camera Pose Regression and Refinement	Samarth Sinha, Jason Y. Zhang, Andrea Tagliasacchi, Igor Gilitschenski, David B. Lindell	University of Toronto	不相关	Camera pose estimation is a key step in standard 3D reconstruction pipelines that operates on a dense set of images of a single object or scene. However, methods for pose estimation often fail when there are only a few images available because they rely on the ability to robustly identify and match visual features between pairs of images. While these methods can work robustly with dense camera views, capturing a large set of images can be time consuming or impractical. Here, we propose Sparse-View Camera Pose Regression and Refinement (SparsePose) for recovering accurate camera poses given a sparse set of wide-baseline images (fewer than 10). The method learns to regress initial camera poses and then iteratively refine them after training on a large-scale dataset of objects (Co3D: Common Objects in 3D). SparsePose significantly outperforms conventional and learning-based baselines in recovering accurate camera rotations and translations. We also demonstrate our pipeline for high-fidelity 3D reconstruction using only 5-9 images of an object.
	27	Proceedings of the IEEE CVPR, 2023, pp.9087-9098	VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion	Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, Anima Anandkumar	NYU	相关(CNN)	Humans can easily imagine the complete 3D geometry of occluded objects and scenes. This appealing ability is vital for recognition and understanding. To enable such capability in AI systems, we propose VoxFormer, a Transformer-based semantic scene completion framework that can output complete 3D volumetric semantics from only 2D images. Our framework adopts a two-stage design where we start from a sparse set of visible and occupied voxel queries from depth estimation, followed by a densification stage that generates dense 3D voxels from the sparse ones. A key idea of this design is that the visual features on 2D images correspond only to the visible scene structures rather than the occluded or empty spaces. Therefore, starting with the featurization and prediction of the visible structures is more reliable. Once we obtain the set of sparse queries, we apply a masked autoencoder design to propagate the information to all the voxels by self-attention. Experiments on SemanticKITTI show that VoxFormer outperforms the state of the art with a relative improvement of 20.0% in geometry and 18.1% in semantics and reduces GPU memory during training to less than 16GB. Our code is available on https://github.com/NVlabs/VoxFormer.
	28	Proceedings of the IEEE CVPR, 2023, pp.21296-21306	Markerless Camera-to-Robot Pose Estimation via Self-Supervised Sim-to-Real Transfer	Jingpei Lu, Florian Richter, Michael C. Yip	University of California, San Diego	不相关	Solving the camera-to-robot pose is a fundamental requirement for vision-based robot control, and is a process that takes considerable effort and cares to make accurate. Traditional approaches require modification of the robot via markers, and subsequent deep learning approaches enabled markerless feature extraction. Mainstream deep learning methods only use synthetic data and rely on Domain Randomization to fill the sim-to-real gap, because acquiring the 3D annotation is labor-intensive. In this work, we go beyond the limitation of 3D annotations for real-world data. We propose an end-to-end pose estimation framework that is capable of online camera-to-robot calibration and a self-supervised training method to scale the training to unlabeled real-world data. Our framework combines deep learning and geometric vision for solving the robot pose, and the pipeline is fully differentiable. To train the Camera-to-Robot Pose Estimation Network (CtRNet), we leverage foreground segmentation and differentiable rendering for image-level self-supervision. The pose prediction is visualized through a renderer and the image loss with the input image is back-propagated to train the neural network. Our experimental results on two public real datasets confirm the effectiveness of our approach over existing works. We also integrate our framework into a visual servoing system to demonstrate the promise of real-time precise robot pose estimation for automation tasks.
	29	Proceedings of the IEEE CVPR, 2023, pp.17990-17999	Event-Guided Person Re-Identification via Sparse-Dense Complementary Learning	Chengzhi Cao, Xueyang Fu, Hongjian Liu, Yukun Huang, Kunyu Wang, Jiebo Luo, Zheng-Jun Zha	University of Science and Technology of China, China	相关(CNN)	Video-based person re-identification (Re-ID) is a prominent computer vision topic due to its wide range of video surveillance applications. Most existing methods utilize spatial and temporal correlations in frame sequences to obtain discriminative person features. However, inevitable degradations, e.g., motion blur contained in frames often cause ambiguity texture noise and temporal disturbance, leading to the loss of identity-discriminating cues. Recently, a new bio-inspired sensor called event camera, which can asynchronously record intensity changes, brings new vitality to the Re-ID task. With the microsecond resolution and low latency, event cameras can accurately capture the movements of pedestrians even in the aforementioned degraded environments. Inspired by the properties of event cameras, in this work, we propose a Sparse-Dense Complementary Learning Framework, which effectively extracts identity features by fully exploiting the complementary information of dense frames and sparse events. Specifically, for frames, we build a CNN-based module to aggregate the dense features of pedestrian appearance step-by-step, while for event streams, we design a bio-inspired spiking neural backbone, which encodes event signals into sparse feature maps in a spiking form, to present the dynamic motion cues of pedestrians. Finally, a cross feature alignment module is constructed to complementarily fuse motion information from events and appearance cues from frames to enhance identity representation learning. Experiments on several benchmarks show that by employing events and SNN into Re-ID, our method significantly outperforms competitive methods.

30	Proceedings of the IEEE CVPR, 2023, pp.18043-18052	Event-Based Frame Interpolation With Ad-Hoc Deblurring	Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, Luc Van Gool	Zhejiang University	不相关	The performance of video frame interpolation is inherently correlated with the ability to handle motion in the input scene. Even though previous works recognize the utility of asynchronous event information for this task, they ignore the fact that motion may or may not result in blur in the input video to be interpolated, depending on the length of the exposure time of the frames and the speed of the motion, and assume either that the input video is sharp, restricting themselves to frame interpolation, or that it is blurry, including an explicit, separate deblurring stage before interpolation in their pipeline. We instead propose a general method for event-based frame interpolation that performs deblurring ad-hoc and thus works both on sharp and blurry input videos. Our model consists in a bidirectional recurrent network that naturally incorporates the temporal dimension of interpolation and fuses information from the input frames and the events adaptively based on their temporal proximity. In addition, we introduce a novel real-world high-resolution dataset with events and color videos which provides a challenging evaluation setting for the examined task. Extensive experiments on the standard GoPro benchmark and on our dataset show that our network consistently outperforms previous state-of-the-art methods on frame interpolation, single image deblurring and the joint task of interpolation and deblurring. Our code and dataset will be available at https://github.com/AHupuJR/REFID.
31	Proceedings of the IEEE CVPR, 2023, pp.13343-13353	X3KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camera 3D Object Detection	Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman Narayanan, Senthil Yogamani, Fatih Porikli	/	相关(LiDAR point cloud)	Recent advances in 3D object detection (3DOD) have obtained remarkably strong results for LiDAR-based models. In contrast, surround-view 3DOD models based on multiple camera images underperform due to the necessary view transformation of features from perspective view (PV) to a 3D world representation which is ambiguous due to missing depth information. This paper introduces X3KD, a comprehensive knowledge distillation framework across different modalities, tasks, and stages for multi-camera 3DOD. Specifically, we propose cross-task distillation from an instance segmentation teacher (X-IS) in the PV feature extraction stage providing supervision without ambiguous error backpropagation through the view transformation. After the transformation, we apply cross-modal feature distillation (X-FD) and adversarial training (X-AT) to improve the 3D world representation of multi-camera features through the information contained in a LiDAR-based 3DOD teacher. Finally, we also employ this teacher for cross-modal output distillation (X-OD), providing dense supervision at the prediction stage. We perform extensive ablations of knowledge distillation at different stages of multi-camera 3DOD. Our final X3KD model outperforms previous state-of-the-art approaches on the nuScenes and Waymo datasets and generalizes to RADAR-based 3DOD. Qualitative results video at https://youtu.be/1do9DPFmr38.
32	Proceedings of the IEEE CVPR, 2023, pp.929-939	NeuMap: Neural Coordinate Mapping by Auto-Transdecoder for Camera Localization	Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, Yasutaka Furukawa	Simon Fraser University	相关	This paper presents an end-to-end neural mapping method for camera localization, dubbed NeuMap, encoding a whole scene into a grid of latent codes, with which a Transformer-based auto-decoder regresses 3D coordinates of query pixels. State-of-the-art feature matching methods require each scene to be stored as a 3D point cloud with per-point features, consuming several gigabytes of storage per scene. While compression is possible, performance drops significantly at high compression rates. Conversely, coordinate regression methods achieve high compression by storing scene information in a neural network but suffer from reduced robustness. NeuMap combines the advantages of both approaches by utilizing 1) learnable latent codes for efficient scene representation and 2) a scene-agnostic Transformer-based auto-decoder to infer coordinates for query pixels. This scene-agnostic network design learns robust matching priors from large-scale data and enables rapid optimization of codes for new scenes while keeping the network weights fixed. Extensive evaluations on five benchmarks show that NeuMap significantly outperforms other coordinate regression methods and achieves comparable performance to feature matching methods while requiring a much smaller scene representation size. For example, NeuMap achieves 39.1% accuracy in the Aachen night benchmark with only 6MB of data, whereas alternative methods require 100MB or several gigabytes and fail completely under high compression settings.
33	Proceedings of the IEEE CVPR, 2023, pp.9275-9285	Depth Estimation From Camera Image and mmWave Radar Point Cloud	Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, Alex Wong	University of California, Los Angeles	相关(LiDAR point cloud)	We present a method for inferring dense depth from a camera image and a sparse noisy radar point cloud. We first describe the mechanics behind mmWave radar point cloud formation and the challenges that it poses, i.e. ambiguous elevation and noisy depth and azimuth components that yields incorrect positions when projected onto the image, and how existing works have overlooked these nuances in camera-radar fusion. Our approach is motivated by these mechanics, leading to the design of a network that maps each radar point to the possible surfaces that it may project onto in the image plane. Unlike existing works, we do not process the raw radar point cloud as an erroneous depth map, but query each raw point independently to associate it with likely pixels in the image -- yielding a semi-dense radar depth map. To fuse radar depth with an image, we propose a gated fusion scheme that accounts for the confidence scores of the correspondence so that we selectively combine radar and camera embeddings to yield a dense depth map. We test our method on the NuScenes benchmark and show a 10.3% improvement in mean absolute error and a 9.1% improvement in root-mean-square error over the best method.
34	Proceedings of the IEEE CVPR, 2023, pp.13924-13934	Learning Event Guided High Dynamic Range Video Reconstruction	Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, Boxin Shi	National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University	不相关	Limited by the trade-off between frame rate and exposure time when capturing moving scenes with conventional cameras, frame based HDR video reconstruction suffers from scene-dependent exposure ratio balancing and ghosting artifacts. Event cameras provide an alternative visual representation with a much higher dynamic range and temporal resolution free from the above issues, which could be an effective guidance for HDR imaging from LDR videos. In this paper, we propose a multimodal learning framework for event guided HDR video reconstruction. In order to better leverage the knowledge of the same scene from the two modalities of visual signals, a multimodal representation alignment strategy to learn a shared latent space and a fusion module tailored to complementing two types of signals for different dynamic ranges in different regions are proposed. Temporal correlations are utilized recurrently to suppress the flickering effects in the reconstructed HDR video. The proposed HDRev-Net demonstrates state-of-the-art performance quantitatively and qualitatively for both synthetic and real-world data.
35	Proceedings of the IEEE CVPR, 2023, pp.4757-4768	Visibility Aware Human-Object Interaction Tracking From Single RGB Camera	Xianghui Xie, Bharat Lal Bhatnagar, Gerard Pons-Moll	University of T¨ubingen, T¨ubingen AI Center, Germany Max Planck Institute for Informatics, Saarland Informatics Campus, Germany	不相关	Capturing the interactions between humans and their environment in 3D is important for many applications in robotics, graphics, and vision. Recent works to reconstruct the 3D human and object from a single RGB image do not have consistent relative translation across frames because they assume a fixed depth. Moreover, their performance drops significantly when the object is occluded. In this work, we propose a novel method to track the 3D human, object, contacts, and relative translation across frames from a single RGB camera, while being robust to heavy occlusions. Our method is built on two key insights. First, we condition our neural field reconstructions for human and object on per-frame SMPL model estimates obtained by pre-fitting SMPL to a video sequence. This improves neural reconstruction accuracy and produces coherent relative translation across frames. Second, human and object motion from visible frames provides valuable information to infer the occluded object. We propose a novel transformer-based neural network that explicitly uses object visibility and human motion to leverage neighboring frames to make predictions for the occluded frames. Building on these insights, our method is able to track both human and object robustly even under occlusions. Experiments on two datasets show that our method significantly improves over the state-of-the-art methods. Our code and pretrained models are available at: https://virtualhumans.mpi-inf.mpg.de/VisTracker.
36	Proceedings of the IEEE CVPR, 2023, pp.13132-13141	Privacy-Preserving Representations Are Not Enough: Recovering Scene Content From Camera Poses	Kunal Chelani, Torsten Sattler, Fredrik Kahl, Zuzana Kukelova	Chalmers University of Technology	不相关	Visual localization is the task of estimating the camera pose from which a given image was taken and is central to several 3D computer vision applications. With the rapid growth in the popularity of AR/VR/MR devices and cloud-based applications, privacy issues are becoming a very important aspect of the localization process. Existing work on privacy-preserving localization aims to defend against an attacker who has access to a cloud-based service. In this paper, we show that an attacker can learn about details of a scene without any access by simply querying a localization service. The attack is based on the observation that modern visual localization algorithms are robust to variations in appearance and geometry. While this is in general a desired property, it also leads to algorithms localizing objects that are similar enough to those present in a scene. An attacker can thus query a server with a large enough set of images of objects, e.g., obtained from the Internet, and some of them will be localized. The attacker can thus learn about object placements from the camera poses returned by the service (which is the minimal information returned by such a service). In this paper, we develop a proof-of-concept version of this attack and demonstrate its practical feasibility. The attack does not place any requirements on the localization algorithm used, and thus also applies to privacy-preserving representations. Current work on privacy-preserving representations alone is thus insufficient.

37	Proceedings of the IEEE CVPR, 2023, pp.8969-8978	Neural Voting Field for Camera-Space 3D Hand Pose Estimation	Lin Huang, Chung-Ching Lin, Kevin Lin, Lin Liang, Lijuan Wang, Junsong Yuan, Zicheng Liu	University at Buffalo	不相关	We present a unified framework for camera-space 3D hand pose estimation from a single RGB image based on 3D implicit representation. As opposed to recent works, most of which first adopt holistic or pixel-level dense regression to obtain relative 3D hand pose and then follow with complex second-stage operations for 3D global root or scale recovery, we propose a novel unified 3D dense regression scheme to estimate camera-space 3D hand pose via dense 3D point-wise voting in camera frustum. Through direct dense modeling in 3D domain inspired by Pixel-aligned Implicit Functions for 3D detailed reconstruction, our proposed Neural Voting Field (NVF) fully models 3D dense local evidence and hand global geometry, helping to alleviate common 2D-to-3D ambiguities. Specifically, for a 3D query point in camera frustum and its pixel-aligned image feature, NVF, represented by a Multi-Layer Perceptron, regresses: (i) its signed distance to the hand surface; (ii) a set of 4D offset vectors (1D voting weight and 3D directional vector to each hand joint). Following a vote-casting scheme, 4D offset vectors from near-surface points are selected to calculate the 3D hand joint coordinates by a weighted average. Experiments demonstrate that NVF outperforms existing state-of-the-art algorithms on FreiHAND dataset for camera-space 3D hand pose estimation. We also adapt NVF to the classic task of root-relative 3D hand pose estimation, for which NVF also obtains state-of-the-art results on HO3D dataset.
38	Proceedings of the IEEE CVPR, 2023, pp.5013-5022	Generating Aligned Pseudo-Supervision From Non-Aligned Data for Image Restoration in Under-Display Camera	Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Jinwei Gu, Chen Change Loy	S-Lab, Nanyang Technological University	不相关	Due to the difficulty in collecting large-scale and perfectly aligned paired training data for Under-Display Camera (UDC) image restoration, previous methods resort to monitor-based image systems or simulation-based methods, sacrificing the realness of the data and introducing domain gaps. In this work, we revisit the classic stereo setup for training data collection -- capturing two images of the same scene with one UDC and one standard camera. The key idea is to "copy" details from a high-quality reference image and "paste" them on the UDC image. While being able to generate real training pairs, this setting is susceptible to spatial misalignment due to perspective and depth of field changes. The problem is further compounded by the large domain discrepancy between the UDC and normal images, which is unique to UDC restoration. In this paper, we mitigate the non-trivial domain discrepancy and spatial misalignment through a novel Transformer-based framework that generates well-aligned yet high-quality target data for the corresponding UDC input. This is made possible through two carefully designed components, namely, the Domain Alignment Module (DAM) and Geometric Alignment Module (GAM), which encourage robust and accurate discovery of correspondence between the UDC and normal views. Extensive experiments show that high-quality and well-aligned pseudo UDC training pairs are beneficial for training a robust restoration network. Code and the dataset are available at
39	Proceedings of the IEEE CVPR, 2023, pp.17366-17375	All-in-Focus Imaging From Event Focal Stack	Hanyue Lou, Minggui Teng, Yixin Yang, Boxin Shi	National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University	不相关	Traditional focal stack methods require multiple shots to capture images focused at different distances of the same scene, which cannot be applied to dynamic scenes well. Generating a high-quality all-in-focus image from a single shot is challenging, due to the highly ill-posed nature of the single-image defocus and deblurring problem. In this paper, to restore an all-in-focus image, we propose the event focal stack which is defined as event streams captured during a continuous focal sweep. Given an RGB image focused at an arbitrary distance, we explore the high temporal resolution of event streams, from which we automatically select refocusing timestamps and reconstruct corresponding refocused images with events to form a focal stack. Guided by the neighbouring events around the selected timestamps, we can merge the focal stack with proper weights and restore a sharp all-in-focus image. Experimental results on both synthetic and real
40	Proceedings of the IEEE CVPR, 2023, pp.1557-1567	Learning Spatial-Temporal Implicit Neural Representations for Event-Guided Video Super-Resolution	Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, Lin Wang	AI Thrust, HKUST(GZ)	不相关	Event cameras sense the intensity changes asynchronously and produce event streams with high dynamic range and low latency. This has inspired research endeavors utilizing events to guide the challenging video super-resolution (VSR) task. In this paper, we make the first at tempt to address a novel problem of achieving VSR at random scales by taking advantages of the high temporal resolution property of events. This is hampered by the difficulties of representing the spatial-temporal information of events when guiding VSR. To this end, we propose a novel framework that incorporates the spatial-temporal interpolation of events to VSR in a unified framework. Our key idea is to learn implicit neural representations from queried spatial-temporal coordinates and features from both RGB frames and events. Our method contains three parts. Specifically, the Spatial-Temporal Fusion (STF) module first learns the 3D features from events and RGB frames. Then, the Temporal Filter (TF) module unlocks more explicit motion information from the events near the queried timestamp and generates the 2D features. Lastly, the Spatial-Temporal Implicit Representation (STIR) module recovers the SR frame in arbitrary resolutions from the outputs of these two modules. In addition, we collect a real-world dataset with spatially aligned events and RGB frames. Extensive experiments show that our method significantly surpass the prior-arts and achieves VSR with random scales, e.g., 6.5. Code and dataset are available at https://.
41	Proceedings of the IEEE CVPR, 2023, pp.21643-21652	MSMDFusion: Fusing LiDAR and Camera at Multiple Scales With Multi-Depth Seeds for 3D Object Detection	Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, Yu-Gang Jiang	Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University	相关(LiDAR)	Fusing LiDAR and camera information is essential for accurate and reliable 3D object detection in autonomous driving systems. This is challenging due to the difficulty of combining multi-granularity geometric and semantic features from two drastically different modalities. Recent approaches aim at exploring the semantic densities of camera features through lifting points in 2D camera images (referred to as "seeds") into 3D space, and then incorporate 2D semantics via cross-modal interaction or fusion techniques. However, depth information is under-investigated in these approaches when lifting points into 3D space, thus 2D semantics can not be reliably fused with 3D points. Moreover, their multi-modal fusion strategy, which is implemented as concatenation or attention, either can not effectively fuse 2D and 3D information or is unable to perform fine-grained interactions in the voxel space. To this end, we propose a novel framework with better utilization of the depth information and fine-grained cross-modal interaction between LiDAR and camera, which consists of two important components. First, a Multi-Depth Unprojection (MDU) method is used to enhance the depth quality of the lifted points at each interaction level. Second, a Gated Modality-Aware Convolution (GMA-Conv) block is applied to modulate voxels involved with the camera modality in a fine-grained manner and then aggregate multi-modal features into a unified space. Together they provide the detection head with more comprehensive features from LiDAR and camera. On the nuScenes test benchmark, our proposed method, abbreviated as MSMDFusion, achieves state-of-the-art results on both 3D object detection and tracking tasks without using test-time-augmentation and ensemble techniques. The code is available at https://github.com/SxJyJay/MSMDFusion.
42	Proceedings of the IEEE CVPR, 2023, pp.21456-21465	Inverting the Imaging Process by Learning an Implicit Camera Model	Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Qing Wang	School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China	不相关	Representing visual signals with implicit coordinate-based neural networks, as an effective replacement of the traditional discrete signal representation, has gained considerable popularity in computer vision and graphics. In contrast to existing implicit neural representations which focus on modelling the scene only, this paper proposes a novel implicit camera model which represents the physical imaging process of a camera as a deep neural network. We demonstrate the power of this new implicit camera model on two inverse imaging tasks: i) generating all-in-focus photos, and ii) HDR imaging. Specifically, we devise an implicit blur generator and an implicit tone mapper to model the aperture and exposure of the camera's imaging process, respectively. Our implicit camera model is jointly learned together with implicit scene models under multi-focus stack and multi-exposure bracket supervision. We have demonstrated the effectiveness of our new model on large number of test images and videos, producing accurate and visually appealing all-in-focus and high dynamic range images. In principle, our new implicit neural camera model has
43	Proceedings of the IEEE CVPR, 2023, pp.1588-1598	Event-Based Blurry Frame Interpolation Under Blind Exposure	Wenming Weng, Yueyi Zhang, Zhiwei Xiong	University of Science and Technology of China, Hefei, China	不相关	Restoring sharp high frame-rate videos from low frame-rate blurry videos is a challenging problem. Existing blurry frame interpolation methods assume a predefined and known exposure time, which suffer from severe performance drop when applied to videos captured in the wild. In this paper, we study the problem of blurry frame interpolation under blind exposure with the assistance of an event camera. The high temporal resolution of the event camera is beneficial to obtain the exposure prior that is lost during the imaging process. Besides, sharp frames can be restored using event streams and blurry frames relying on the mutual constraint among them. Therefore, we first propose an exposure estimation strategy guided by event streams to estimate the lost exposure prior, transforming the blind exposure problem well-posed. Second, we propose to model the mutual constraint with a temporal-exposure control strategy through iterative residual learning. Our blurry frame interpolation method achieves a distinct performance boost over existing methods on both synthetic and self-collected real-world datasets under blind exposure.

	44	Proceedings of the IEEE CVPR, 2023, pp.5642-5651	Data-Driven Feature Tracking for Event Cameras	Nico Messikommer, Carter Fang, Mathias Gehrig, Davide Scaramuzza	Robotics and Perception Group, University of Zurich, Switzerland	不相关	Because of their high temporal resolution, increased resilience to motion blur, and very sparse output, event cameras have been shown to be ideal for low-latency and low-bandwidth feature tracking, even in challenging scenarios. Existing feature tracking methods for event cameras are either handcrafted or derived from first principles but require extensive parameter tuning, are sensitive to noise, and do not generalize to different scenarios due to unmodeled effects. To tackle these deficiencies, we introduce the first data-driven feature tracker for event cameras, which leverages low-latency events to track features detected in a grayscale frame. We achieve robust performance via a novel frame attention module, which shares information across feature tracks. By directly transferring zero-shot from synthetic to real data, our data-driven tracker outperforms existing approaches in relative feature age by up to 120% while also achieving the lowest latency. This performance gap is further increased to 130% by adapting our tracker to real data with a novel self-supervision strategy.
	45	Proceedings of the IEEE CVPR, 2023, pp.22180-22190	1000 FPS HDR Video With a Spike-RGB Hybrid Camera	Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, Boxin Shi	National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University	不相关	Capturing high frame rate and high dynamic range (HFR&HDR) color videos in high-speed scenes with conventional frame-based cameras is very challenging. The increasing frame rate is usually guaranteed by using shorter exposure time so that the captured video is severely interfered by noise. Alternating exposures could alleviate the noise issue but sacrifice frame rate due to involving long-exposure frames. The neuromorphic spiking camera records high-speed scenes of high dynamic range without colors using a completely different sensing mechanism and visual representation. We introduce a hybrid camera system composed of a spiking and an alternating-exposure RGB camera to capture HFR&HDR scenes with high fidelity. Our insight is to bring each camera's superiority into full play. The spike frames, with accurate fast motion information encoded, are first reconstructed for motion representation, from which the spike-based optical flows guide the recovery of missing temporal information for middle- and long-exposure RGB images while retaining their reliable color appearances. With the strong temporal constraint estimated from spike trains, both missing and distorted colors cross RGB frames are recovered to generate time-consistent and HFR color frames. We collect a new Spike-RGB dataset that contains 300 sequences of synthetic data and 20 groups of real-world data to demonstrate 1000 FPS HDR videos outperforming HDR video reconstruction methods and commercial high-speed cameras.

Event camera

发表年限	序号	会议/期刊名	论文题目	论文作者	单位	论文方法是否与图神经/深度/卷积网络/激光雷达/点云相	论文摘要
	1	Proceedings of the IEEE ICCV, 2017, pp. 1-10	Globally-Optimal Inlier Set Maximisation for Simultaneous Camera Pose and Feature Correspondence	Dylan Campbell, Lars Petersson, Laurent Kneip, Hongdong Li	Australian National University	不相关	Estimating the 6-DoF pose of a camera from a single image relative to a pre-computed 3D point-set is an important task for many computer vision applications. Perspective-n-Point (PnP) solvers are routinely used for camera pose estimation, provided that a good quality set of 2D-3D feature correspondences are known beforehand. However, finding optimal corresponden-ces between 2D key-points and a 3D point-set is non-trivial, especially when only geometric (position) information is known. Existing approaches to the simultaneous pose and corresponde-nce problem use local optimisation, and are therefore unlikely to find the optimal solution without a good pose initialisation, or introduce restrictive assumptions. Since a large proportion of outliers are common for this problem, we instead propose a globally-optimal inlier set cardinality maximisation approach which jointly estimates optimal camera pose and optimal correspon-dences. Our approach employs branch-and-bound to search the 6D space of camera poses, guaranteeing global optimality without requiring a pose prior. The geometry of SE(3) is used to find novel upper and lower bounds for the number of inliers and local optimisation is integrated to accelerate convergence. The evaluation empirically supports the optimality proof and shows that the method performs much more robustly than existing approaches, including on a
	2	Proceedings of the IEEE ICCV, 2017, pp. 29-38	Distributed Very Large Scale Bundle Adjustment by Global Camera Consensus	Runze Zhang, Siyu Zhu, Tian Fang, Long Quan	Department of Computer Science and Engineering The Hong Kong University of Science and Technology	不相关	The increasing scale of Structure-from-Motion is fundamentally limited by the conventional optimization framework for the all-in-one global bundle adjustment. In this paper, we propose a distributed approach to coping with this global bundle adjustment for very large scale Structure-from-Motion computation. First, we derive the distributed formulation from the classical optimization algorithm ADMM, Alternating Direction Method of Multipliers, based on the global camera consensus. Then, we analyze the conditions under which the convergence of this distributed optimization would be guaranteed. In particular, we adopt over-relaxation and self-adaption schemes to improve the convergence rate. After that, we propose to split the large scale camera-point visibility graph in order to reduce the communication overheads of the distributed computing. The experiments on both public large scale SfM data-sets and our very large scale aerial photo sets demonstrate that the proposed distributed method clearly outperforms the state-of-the-art method in
	3	Proceedings of the IEEE ICCV, 2017, pp. 271-279	Deep Occlusion Reasoning for Multi-Camera Multi-Target Detection	Pierre Baque, Francois Fleuret, Pascal Fua	CVLab, EPFL, Lausanne, Switzerland	相关(CNN)	People detection in 2D images has improved greatly in recent years. However, comparatively little of this progress has percolated into multi-camera multi-people tracking algorithms, whose performance still degrades severely when scenes become very crowded. In this work, we introduce a new architecture that combines Convolutional Neural Nets and Conditional Random Fields to explicitly resolve ambiguities. One of its key ingredients are high-order CRF terms that model potential occlusions and give our approach its robustness even when many people are present. Our model is trained end-to-end and we show that it outperforms several state-of-the-art algorithms on challenging scenes.
	4	Proceedings of the IEEE ICCV, 2017, pp. 736-744	Complex Event Detection by Identifying Reliable Shots From Untrimmed Videos	Hehe Fan, Xiaojun Chang, De Cheng, Yi Yang, Dong Xu, Alexander G. Hauptmann	Centre for Artificial Intelligence, University of Technology Sydney, Sydney, Australia	不相关	The goal of complex event detection is to automatically detect whether an event of interest happens in temporally untrimmed long videos which usually consist of multiple video shots. Observing some video shots in positive (resp. negative) videos are irrelevant (resp. relevant) to the given event class, we formulate this task as a multi-instance learning (MIL) problem by taking each video as a bag and the video shots in each video as instances. To this end, we propose a new MIL method, which simultaneously learns a linear SVM classifier and infers a binary indicator for each instance in order to select reliable training instances from each positive or negative bag. In our new objective function, we balance the weighted training errors and a l1-l2 mixed-norm regularization term which adaptively selects reliable shots as training instances from different videos to have them as diverse as possible. We also develop an alternating optimization approach that can efficiently solve our proposed objective function. Extensive experiments on the challenging real-world Multimedia Event Detection (MED) datasets MEDTest-14, MEDTest-13 and CCV clearly demonstrate the effectiveness of our proposed MIL approach for complex event
	5	Proceedings of the IEEE ICCV, 2017, pp.910-919	BodyFusion: Real-Time Capture of Human Motion and Surface Geometry Using a Single Depth Camera	Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, Yebin Liu	Beihang University, Beijing, China	不相关	We propose BodyFusion, a novel real-time geometry fusion method that can track and reconstruct non-rigid surface motion of a human performance using a single consumer-grade depth camera. To reduce the ambiguities of the non-rigid deformation parameterization on the surface graph nodes, we take advantage of the internal articulated motion prior for human performance and contribute a skeleton-embedded surface fusion (SSF) method. The key feature of our method is that it jointly solves for both the skeleton and graph-node deformations based on information of the attachments between the skeleton and the graph nodes. The attachments are also updated frame by frame based on the fused surface geometry and the computed deformations. Overall, our method enables increasingly denoised, detailed, and complete surface reconstruction as well as the updating of the skeleton and attachments as the temporal depth frames are fused. Experimental results show that our method exhibits substantially improved nonrigid motion fusion performance and tracking robustness compared with previous state-of-the-art fusion methods. We also contribute a dataset for the quantitative evaluation of fusion-based dynamic scene
	6	Proceedings of the IEEE ICCV, 2017, pp.2372-2381	Efficient Global 2D-3D Matching for Camera Localization in a Large-Scale 3D Map	Liu Liu, Hongdong Li, Yuchao Dai	Northwestern Polytechnical University, Xi'an, China	不相关	Given an image of a street scene in a city, this paper develops a new method that can quickly and precisely pinpoint at which location (as well as viewing direction) the image was taken, against a pre-stored large-scale 3D point-cloud map of the city. We adopt the recently developed 2D-3D direct feature matching framework for this task [23,31,32,42-44]. This is a challenging task especially for large-scale problems. As the map size grows bigger, many 3D points in the wider geographical area can be visually very similar-or even identical-causing severe ambiguities in 2D-3D feature matching. The key is to quickly and unambiguously find the correct matches between a query image and the large 3D map. Existing methods solve this problem mainly via comparing individual features' visual similarities in a local and per feature manner, thus only local solutions can be found, inadequate for large-scale applications. In this paper, we introduce a global method which harnesses global contextual information exhibited both within the query image and among all the 3D points in the map. This is achieved by a novel global ranking algorithm, applied to a Markov network built upon the 3D map, which takes account of not only visual similarities between individual 2D-3D matches, but also their global compatibilities (as measured by co-visibility) among all matching pairs found in the scene. Tests on standard benchmark datasets show that our method achieved both higher precision and comparable recall, compared with the state-of-the-art.

2017	7	Proceedings of the IEEE ICCV, 2017, pp.3647-3656	Leveraging Weak Semantic Relevance for Complex Video Event Classification	Chao Li, Jiewei Cao, Zi Huang, Lei Zhu, Heng Tao Shen	The University of Queensland Australia	不相关	Existing video event classification approaches suffer from limited human-labeled semantic annotations. Weak semantic annotations can be harvested from Web-knowledge without involving any human interaction. However such weak annotations are noisy, thus can not be effectively utilized without distinguishing its reliability. In this paper, we propose a novel approach to automatically maximize the utility of weak semantic annotations (formalized as the semantic relevance of video shots to the target event) to facilitate video event classification. A novel attention model is designed to determine the attention scores of video shots, where the weak semantic relevance is considered as attentional guidance. Specifically, our model jointly optimizes two objectives at different levels. The first one is the classification loss corresponding to video-level groundtruth labels, and the second is the shot-level relevance loss corresponding to weak semantic relevance. We use a long short-term memory (LSTM) layer to capture the temporal information carried by the shots of a video. In each timestep, the LSTM employs the attention model to weight the current shot under the guidance of its weak semantic relevance to the event of interest. Thus, we can automatically exploit weak semantic relevance to assist video event classification. Extensive experiments have been conducted on three complex large-scale video event datasets i.e., MEDTest14, ActivityNet and FCVID. Our approach achieves the state-of-the-art classification performance on all three datasets. The significant performance improvement upon the conventional attention model also demonstrates the effectiveness of our model.
	8	Proceedings of the IEEE ICCV, 2017, pp.4326-4334	Summarization and Classification of Wearable Camera Streams by Learning the Distributions Over Deep Features of Out-Of-Sample Image Sequences	Alessandro Perina, Sadegh Mohammadi, Nebojsa Jojic, Vittorio Murino	WDG Core Data - Microsoft Corp	相关(CNN)	A popular approach to training classifiers of new image classes is to use lower levels of a pre-trained feed-forward neural network and retrain only the top. Thus, most layers simply serve as highly nonlinear feature extractors. While these features were found useful for classifying a variety of scenes and objects, previous work also demonstrated unusual levels of sensitivity to the input especially for images which are veering too far away from the training distribution. This can lead to surprising results as an imperceptible change in an image can be enough to completely change the predicted class. This occurs in particular in applications involving personaldata, typically acquired with wearable cameras (e.g., visual lifelogs), where the problem is also made more complex by the dearth of new labeled training data that make supervised learning with deep models difficult. To alleviate these problems, in this paper we propose a new generative model that captures the feature distribution in new data. Its latent space then becomes more representative of the new data, while still retaining the generalization properties. In particular, we use constrained Markov walks over a counting grid for modeling image sequences, which not only yield good latent representations, but allow for excellent classification with only a handful of labeled training examples of the new scenes or objects, a scenario typical in
	9	Proceedings of the IEEE ICCV, 2017, pp.4613-4621	Joint Estimation of Camera Pose, Depth, Deblurring, and Super-Resolution From a Blurred Image Sequence	Haesol Park, Kyoung Mu Lee	Department of ECE, ASRI, Seoul National University, 151-742, Seoul, Korea	不相关	The conventional methods for estimating camera poses and scene structures from severely blurry or low resolution images often result in failure. The off-the-shelf deblurring or super resolution methods may show visually pleasing results. However, applying each technique independently before matching is generally unprofitable because this naive series of procedures ignores the consistency between images. In this paper, we propose a pioneering unified framework that solves four problems simultaneously, namely, dense depth reconstruction, camera pose estimation, super resolution, and deblurring. By reflecting a physical imaging process, we formulate a cost minimization problem and solve it using an alternating optimization technique. The experimental results on both synthetic and real videos show high-quality depth maps derived from severely degraded images that contrast the failures of naive multi-view stereo methods. Our proposed method also produces outstanding deblurred and super-resolved images unlike the independent application or combination of conventional video deblurring, super resolution methods.
	10	Proceedings of the IEEE ICCV, 2017, pp.4668-4676	Dynamics Enhanced Multi-Camera Motion Segmentation From Unsynchronized Videos	Xikang Zhang, Bengisu Ozbay, Mario Sznaier, Octavia Camps	Electrical and Computer Engineering Northeastern University, Boston MA 02115, US	不相关	This paper considers the multi-camera motion segmentation problem using unsynchronized videos. Specifically, given two video clips containing several moving objects, captured by unregistered, unsynchronized cameras with different viewpoints, our goal is to assign features to moving objects in the scene. This problem challenges existing methods, due to the lack of registration information and correspondences across cameras. To solve it, we propose a new method that exploits both shape and dynamical information and does not require spatio-temporal registration or shared features. As shown in the paper, the combination of shape and dynamical information results in improved performance even in the single camera case, and allows for solving the multi-camera segmentation problem with a computational cost similar to that of existing single-view techniques. These results are illustrated using both the existing Hopkins 155 data set and a new multi-camera data set, the RSL-12.
	11	Proceedings of the IEEE ICCV, 2017, pp.5170-5178	What Is Around the Camera?	Stamatios Georgoulis, Konstantinos Rematas, Tobias Ritschel, Mario Fritz, Tinne Tuytelaars, Luc Van Gool	KU Leuven	不相关	How much does a single image reveal about the environment it was taken in? In this paper, we investigate how much of that information can be retrieved from a foreground object, combined with the background (i.e. the visible part of the environment). Assuming it is not perfectly diffuse, the foreground object acts as a complexly shaped and far-from-perfect mirror. An additional challenge is that its appearance confounds the light coming from the environment with the unknown materials it is made of. We propose a learning-based approach to predict the environment from multiple reflectance maps that are computed from approximate surface normals. The proposed method allows us to jointly model the statistics of environments and material properties. We train our system from synthesized training data, but demonstrate its applicability to real-world data. Interestingly, our analysis shows that the information obtained from objects made out of multiple materials often is complementary and leads to better performance.
	12	Proceedings of the IEEE ICCV, 2017, pp.5334-5343	Camera Calibration by Global Constraints on the Motion of Silhouettes	Gil Ben-Artzi	Weizmann Institute of Science Rehovot, Israel	不相关	We address the problem of epipolar geometry using the motion of silhouettes. Such methods match epipolar lines or frontier points across views, which are then used as the set of putative correspondences. We introduce an approach that improves by two orders of magnitude the performance over state-of-the-art methods, by significantly reducing the number of outliers in the putative matching. We model the frontier points' correspondence problem as constrained flow optimization, requiring small differences between their coordinates over consecutive frames. Our approach is formulated as a Linear Integer Program and we show that due to the nature of our problem, it can be solved efficiently in an iterative manner. Our method was validated on four standard datasets providing accurate calibrations across very different viewpoints.
	13	Proceedings of the IEEE ICCV, 2017, pp.5344-5352	Deltille Grids for Geometric Camera Calibration	Hyowon Ha, Michal Perdoch, Hatem Alismail, In So Kweon, Yaser Sheikh	Korea Advanced Institute of Science and Technology	不相关	The recent proliferation of high resolution cameras presents an opportunity to achieve unprecedented levels of precision in visual 3D reconstruction. Yet the camera calibration pipeline, developed decades ago using checkerboards, has remained the de facto standard. In this paper, we ask the question: are checkerboards the optimal pattern for high precision calibration? We empirically demonstrate that deltille grids (regular triangular tiling) produce the highest precision calibration of the possible tilings of Euclidean plane. We posit that they should be the new standard for high-precision calibration and present a complete ecosystem for calibration using deltille grids including: (1) a highly precise corner detection algorithm based on polynomial surface fitting; (2) an indexing scheme based on polarities extracted from the fitted surfaces; and (3) a 2D coding system for deltille grids, which we refer to as DelTags, in lieu of conventional matrix barcodes. We demonstrate state-of-the-art performance and apply the full calibration ecosystem through the use of 3D calibration objects for multiview camera calibration.

	14	Proceedings of the IEEE ICCV, 2017, pp.5353-5361	A Lightweight Single-Camera Polarization Compass With Covariance Estimation	Wolfgang Sturzl	Institute of Robotics and Mechatronics, German Aerospace Center (DLR)	不相关	A lightweight visual compass system is presented as well as a direct method for estimating sun direction and its covariance. The optical elements of the system are described enabling estimation of sky polarization in a FOV of approx. 56 degrees with a single standard camera sensor. Using the proposed direct method, the sun direction and its covariance matrix can be estimated based on the polarization measured in the image plane. Experiments prove the applicability of the polarization sensor and the proposed estimation method, even in difficult conditions. It is also shown that in case the sensor is not leveled, combination with an IMU allows to determine all degrees of orientation. Due to the low weight of the sensor and the low complexity of the estimation method the polarization system is well suited for MAVs which have limited payload and computational resources. Furthermore, since not just the sun direction but also its covariance is estimated an integration in a multi-sensor navigation framework is straight forward.
	1	Proceedings of the IEEE ICCV, 2019, pp.42-51	SANet: Scene Agnostic Network for Camera Localization	Luwei Yang, Ziqian Bai, Chengzhou Tang, Honghua Li, Yasutaka Furukawa, Ping Tan	Simon Fraser University	相关(CNN)	This paper presents a scene agnostic neural architecture for camera localization, where model parameters and scenes are independent from each other.Despite recent advancement in learning based methods, most approaches require training for each scene one by one, not applicable for online applications such as SLAM and robotic navigation, where a model must be built on-the-fly.Our approach learns to build a hierarchical scene representation and predicts a dense scene coordinate map of a query RGB image on-the-fly given an arbitrary scene. The 6D camera pose of the query image can be estimated with the predicted scene coordinate map. Additionally, the dense prediction can be used for other online robotic and AR applications such as obstacle avoidance. We demonstrate the effectiveness and efficiency of our method on both indoor and outdoor benchmarks, achieving state-of-
	2	Proceedings of the IEEE ICCV, 2019, pp.1082-1090	QUARCH: A New Quasi-Affine Reconstruction Stratum From Vague Relative Camera Orientation Knowledge	Devesh Adlakha, Adlane Habed, Fabio Morbidi, Cedric Demonceaux, Michel de Mathelin	ICube laboratory, CNRS, University of Strasbourg	不相关	We present a new quasi-affine reconstruction of a scene and its application to camera self-calibration. We refer to this reconstruction as QUARCH (QUasi-Affine Reconstruction with respect to Camera centers and the Hodographs of horopters). A QUARCH can be obtained by solving a semidefinite programming problem when, (i) the images have been captured by a moving camera with constant intrinsic parameters, and (ii) a vague knowledge of the relative orientation (under or over 120 degrees) between camera pairs is available. The resulting reconstruction comes close enough to an affine one allowing thus an easy upgrade of the QUARCH to its affine and metric counterparts. We also present a constrained Levenberg-Marquardt method for nonlinear optimization subject to Linear Matrix Inequality (LMI) constraints so as to ensure that the QUARCH LMIs are satisfied during optimization. Experiments with synthetic and real data show the benefits of QUARCH in reliably obtaining a metric reconstruction.
	3	Proceedings of the IEEE ICCV, 2019, pp.1335-1344	A Camera That CNNs: Towards Embedded Neural Networks on Pixel Processor Arrays	Laurie Bose, Jianing Chen, Stephen J. Carey, Piotr Dudek, Walterio Mayol-Cuevas	University of Bristol, Bristol, United Kingdom	相关(CNN)	We present a convolutional neural network implementation for pixel processor array (PPA) sensors. PPA hardware consists of a fine-grained array of general-purpose processing elements, each capable of light capture, data storage, program execution, and communication with neighboring elements. This allows images to be stored and manipulated directly at the point of light capture, rather than having to transfer images to external processing hardware. Our CNN approach divides this array up into 4x4 blocks of processing elements, essentially trading-off image resolution for increased local memory capacity per 4x4 "pixel". We implement parallel operations for image addition, subtraction and bit-shifting images in this 4x4 block format. Using these components we formulate how to perform ternary weight convolutions upon these images, compactly store results of such convolutions, perform max-pooling, and transfer the resulting sub-sampled data to an attached micro-controller. We train ternary weight filter CNNs for digit recognition and a simple tracking task, and demonstrate inference of these networks upon the SCAMP5 PPA system. This work represents a first step towards embedding neural network processing
	4	Proceedings of the IEEE ICCV, 2019, pp.1486-1496	Privacy Preserving Image Queries for Camera Localization	Pablo Speciale, Johannes L. Schonberger, Sudipta N. Sinha, Marc Pollefeys	ETH Z¨urich	相关(CNN/point cloud)	Augmented/mixed reality and robotic applications are increasingly relying on cloud-based localization services, which require users to upload query images to perform camera pose estimation on a server. This raises significant privacy concerns when consumers use such services in their homes or in confidential industrial settings. Even if only image features are uploaded, the privacy concerns remain as the images can be reconstructed fairly well from feature locations and descriptors. We propose to conceal the content of the query images from an adversary on the server or a man-in-the-middle intruder. The key insight is to replace the 2D image feature points in the query image with randomly oriented 2D lines passing through their original 2D positions. It will be shown that this feature representation hides the image contents, and thereby protects user privacy, yet still provides sufficient geometric constraints to enable robust and accurate 6-DOF camera pose estimation from feature correspondences. Our proposed method can handle single- and multi-image queries as well as exploit additional information about known structure, gravity, and scale. Numerous experiments demonstrate the high practical relevance of our approach.
	5	Proceedings of the IEEE ICCV, 2019, pp.1497-1505	Calibration Wizard: A Guidance System for Camera Calibration Based on Modelling Geometric and Corner Uncertainty	Songyou Peng, Peter Sturm	ETH Zurich	不相关	It is well known that the accuracy of a calibration depends strongly on the choice of camera poses from which images of a calibration object are acquired. We present a system -- Calibration Wizard -- that interactively guides a user towards taking optimal calibration images. For each new image to be taken, the system computes, from all previously acquired images, the pose that leads to the globally maximum reduction of expected uncertainty on intrinsic parameters and then guides the user towards that pose. We also show how to incorporate uncertainty in corner point position in a novel principled manner, for both, calibration and computation of the next best pose. Synthetic and real-world experiments are performed to demonstrate the effectiveness of Calibration Wizard.
	6	Proceedings of the IEEE ICCV, 2019, pp.1527-1537	Learning an Event Sequence Embedding for Dense Event-Based Deep Stereo	Stepan Tulyakov, Francois Fleuret, Martin Kiefel, Peter Gehler, Michael Hirsch	Space Engineering Center at Ecole´ Polytechnique F´ed´erale de Lausanne	相关	Today, a frame-based camera is the sensor of choice for machine vision applications. However, these cameras, originally developed for acquisition of static images rather than for sensing of dynamic uncontrolled visual environments, suffer from high power consumption, data rate, latency and low dynamic range. An event-based image sensor addresses these drawbacks by mimicking a biological retina. Instead of measuring the intensity of every pixel in a fixed time-interval, it reports events of significant pixel intensity changes. Every such event is represented by its position, sign of change, and timestamp, accurate to the microsecond. Asynchronous event sequences require special handling, since traditional algorithms work only with synchronous, spatially gridded data. To address this problem we introduce a new module for event sequence embedding, for use in difference applications. The module builds a representation of an event sequence by firstly aggregating information locally across time, using a novel fully-connected layer for an irregularly sampled continuous domain, and then across discrete spatial domain. Based on this module, we design a deep learning-based stereo method for event-based cameras. The proposed method is the first learning-based stereo method for an event-based camera and the only method that produces dense results. We show that large performance increases on the Multi Vehicle Stereo Event Camera Dataset (MVSEC), which became the standard set for benchmarking of event-based stereo methods.

7	Proceedings of the IEEE ICCV, 2019, pp.2414-2423	Pro-Cam SSfM: Projector-Camera System for Structure and Spectral Reflectance From Motion	Chunyu Li, Yusuke Monno, Hironori Hidaka, Masatoshi Okutomi	Tokyo Institute of Technology, Tokyo, Japan	相关(point cloud)	In this paper, we propose a novel projector-camera system for practical and low-cost acquisition of a dense object 3D model with the spectral reflectance property. In our system, we use a standard RGB camera and leverage an off-the-shelf projector as active illumination for both the 3D reconstruction and the spectral reflectance estimation. We first reconstruct the 3D points while estimating the poses of the camera and the projector, which are alternately moved around the object, by combining multi-view structured light and structure-from-motion (SfM) techniques. We then exploit the projector for multispectral imaging and estimate the spectral reflectance of each 3D point based on a novel spectral reflectance estimation model considering the geometric relationship between the reconstructed 3D points and the estimated projector positions. Experimental results on several real objects demonstrate that our system can precisely acquire a dense 3D model with the full spectral reflectance property using off-the-shelf devices.
8	Proceedings of the IEEE ICCV, 2019, pp.2841-2850	Local Supports Global: Deep Camera Relocalization With Sequence Enhancement	Fei Xue, Xin Wang, Zike Yan, Qiuyuan Wang, Junqiu Wang, Hongbin Zha	UISEE Technology Inc	相关(CNN)	We propose to leverage the local information in a image sequence to support global camera relocalization. In contrast to previous methods that regress global poses from single images, we exploit the spatial-temporal consistency in sequential images to alleviate uncertainty due to visual ambiguities by incorporating a visual odometry (VO) component. Specifically, we introduce two effective steps called content-augmented pose estimation and motion-based refinement. The content-augmentation step focuses on alleviating the uncertainty of pose estimation by augmenting the observation based on the co-visibility in local maps built by the VO stream. Besides, the motion-based refinement is formulated as a pose graph, where the camera poses are further optimized by adopting relative poses provided by the VO component as additional motion constraints. Thus, the global consistency can be guaranteed. Experiments on the public indoor 7-Scenes and outdoor Oxford RobotCar benchmark datasets demonstrate that benefited from local information inherent in the sequence, our approach outperforms state-of-the-art methods, especially in some challenging cases, e.g., insufficient texture, highly repetitive textures, similar
9	Proceedings of the IEEE ICCV, 2019, pp.2871-2880	CamNet: Coarse-to-Fine Retrieval for Camera Re-Localization	Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, Ping Luo	The University of Hong Kong	相关(CNN)	Camera re-localization is an important but challenging task in applications like robotics and autonomous driving. Recently, retrieval-based methods have been considered as a promising direction as they can be easily generalized to novel scenes. Despite significant progress has been made, we observe that the performance bottleneck of previous methods actually lies in the retrieval module. These methods use the same features for both retrieval and relative pose regression tasks which have potential conflicts in learning. To this end, here we present a coarse-to-fine retrieval-based deep learning framework, which includes three steps, i.e., image-based coarse retrieval, pose-based fine retrieval and precise relative pose regression. With our carefully designed retrieval module, the relative pose regression task can be surprisingly simpler. We design novel retrieval losses with batch hard sampling criterion and two-stage retrieval to locate samples that adapt to the relative pose regression task. Extensive experiments show that our model (CamNet) outperforms the state-of-the-art methods by a large margin on both indoor and outdoor datasets.
10	Proceedings of the IEEE ICCV, 2019, pp. 4111-4119	Enhancing Low Light Videos by Exploring High Sensitivity Camera Noise	Wei Wang, Xin Chen, Cheng Yang, Xiang Li, Xuemei Hu, Tao Yue	Nanjing University, Nanjing, China	不相关	Enhancing low light videos, which consists of denoising and brightness adjustment, is an intriguing but knotty problem. Under low light condition, due to high sensitivity camera setting, commonly negligible noises become obvious and severely deteriorate the captured videos. To recover high quality videos, a mass of image/video denoising/enhancing algorithms are proposed, most of which follow a set of simple assumptions about the statistic characters of camera noise, e.g., independent and identically distributed(i.i.d.), white, additive, Gaussian, Poisson or mixture noises. However, the practical noise under high sensitivity setting in real captured videos is complex and inaccurate to model with these assumptions. In this paper, we explore the physical origins of the practical high sensitivity noise in digital cameras, model them mathematically, and propose to enhance the low light videos based on the noise model by using an LSTM-based neural network. Specifically, we generate the training data with the proposed noise model and train the network with the dark noisy video as input and clear-bright video as output. Extensive comparisons on both synthetic and real captured low light videos with the state-of-the-art methods are conducted to demonstrate the effectiveness of the proposed method.
11	Proceedings of the IEEE ICCV, 2019, pp.5633-5643	End-to-End Learning of Representations for Asynchronous Event-Based Data	Daniel Gehrig, Antonio Loquercio, Konstantinos G. Derpanis, Davide Scaramuzza	Robotics and Perception Group Depts. Informatics and Neuroinformatics University of Zurich and ETH Zurich	不相关	Event cameras are vision sensors that record asynchronous streams of per-pixel brightness changes, referred to as "events". They have appealing advantages over frame based cameras for computer vision, including high temporal resolution, high dynamic range, and no motion blur. Due to the sparse, non-uniform spatio-temporal layout of the event signal, pattern recognition algorithms typically aggregate events into a grid-based representation and subsequently process it by a standard vision pipeline, e.g., Convolutional Neural Network (CNN). In this work, we introduce a general framework to convert event streams into grid-based representations by means of strictly differentiable operations. Our framework comes with two main advantages: (i) allows learning the input event representation together with the task dedicated network in an end to end manner, and (ii) lays out a taxonomy that unifies the majority of extant event representations in the literature and identifies novel ones. Empirically, we show that our approach to learning the event representation end-to-end yields an improvement of approximately 12% on optical flow estimation and object recognition over state-of-the-art methods.
12	Proceedings of the IEEE ICCV, 2019, pp.5664-5673	3D Scene Graph: A Structure for Unified Semantics, 3D Space, and Camera	Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, Silvio Savarese	Stanford University	相关(CNN)	A comprehensive semantic understanding of a scene is important for many applications - but in what space should diverse semantic information (e.g., objects, scene categories, material types, 3D shapes, etc.) be grounded and what should be its structure? Aspiring to have one unified structure that hosts diverse types of semantics, we follow the Scene Graph paradigm in 3D, generating a 3D Scene Graph. Given a 3D mesh and registered panoramic images, we construct a graph that spans the entire building and includes semantics on objects (e.g., class, material, shape and other attributes), rooms (e.g., function, illumination type, etc.) and cameras (e.g., location, etc.), as well as the relationships among these entities. However, this process is prohibitively labor heavy if done manually. To alleviate this we devise a semi-automatic framework that employs existing detection methods and enhances them using two main constraints: I. framing of query images sampled on panoramas to maximize the performance of 2D detectors, and II. multi-view consistency enforcement across 2D detections that originate in different camera locations.
13	Proceedings of the IEEE ICCV, 2019, pp.6292-6300	Dual Attention Matching for Audio-Visual Event Localization	Yu Wu, Linchao Zhu, Yan Yan, Yi Yang	Baidu Research	不相关	In this paper, we investigate the audio-visual event localization problem. This task is to localize a visible and audible event in a video. Previous methods first divide a video into short segments, and then fuse visual and acoustic features at the segment level. The duration of these segments is usually short, making the visual and acoustic feature of each segment possibly not well aligned. Direct concatenation of the two features at the segment level can be vulnerable to a minor temporal misalignment of the two signals. We propose a Dual Attention Matching (DAM) module to cover a longer video duration for better high-level event information modeling, while the local temporal information is attained by the global cross-check mechanism. Our premise is that one should watch the whole video to understand the high-level event, while shorter segments should be checked in detail for localization. Specifically, the global feature of one modality queries the local feature in the other modality in a bi-directional way. With temporal co-occurrence encoded between auditory and visual signals, DAM can be readily applied in various audio-visual event localization tasks, e.g., cross-modality localization, supervised event localization. Experiments on the AVE dataset show our method outperforms the state-of-the-art by a large margin.

2019	14	Proceedings of the IEEE ICCV, 2019, pp.6922-6931	Unsupervised Person Re-Identification by Camera-Aware Similarity Consistency Learning	Ancong Wu, Wei-Shi Zheng, Jian-Huang Lai	School of Electronics and Information Technology, Sun Yat-sen University, China	不相关	For matching pedestrians across disjoint camera views in surveillance, person re-identification (Re-ID) has made great progress in supervised learning. However, it is infeasible to label data in a number of new scenes when extending a Re-ID system. Thus, studying unsupervised learning for Re-ID is important for saving labelling cost. Yet, cross-camera scene variation is a key challenge for unsupervised Re-ID, such as illumination, background and viewpoint variations, which cause domain shift in the feature space and result in inconsistent pairwise similarity distributions that degrade matching performance. To alleviate the effect of cross-camera scene variation, we propose a Camera-Aware Similarity Consistency Loss to learn consistent pairwise similarity distributions for intra-camera matching and cross-camera matching. To avoid learning ineffective knowledge in consistency learning, we preserve the prior common knowledge of intra-camera matching in the pretrained model as reliable guiding information, which does not suffer from cross-camera scene variation as cross-camera matching. To learn similarity consistency more effectively, we further develop a coarse-to-fine consistency learning scheme to learn consistency globally and locally in two steps. Experiments show that our method outperformed the state-of-the-art unsupervised Re-ID methods.
	15	Proceedings of the IEEE ICCV, 2019, pp.7063-7072	Self-Supervised Learning With Geometric Constraints in Monocular Video: Connecting Flow, Depth, and Camera	Yuhua Chen, Cordelia Schmid, Cristian Sminchisescu	Google Research	不相关	We present GLNet, a self-supervised framework for learning depth, optical flow, camera pose and intrinsic parameters from monocular video -- addressing the difficulty of acquiring realistic ground-truth for such tasks. We propose three contributions: 1) we design new loss functions that capture multiple geometric constraints (eg. epipolar geometry) as well as adaptive photometric loss that supports multiple moving objects, rigid and non-rigid, 2) we extend the model such that it predicts camera intrinsics, making it applicable to uncalibrated video, and 3) we propose several online refinement strategies that rely on the symmetry of our self-supervised loss in training and testing, in particular optimizing model parameters and/or the output of different tasks, leveraging their mutual interactions. The idea of jointly optimizing the system output, under all geometric and photometric constraints can be viewed as a dense generalization of classical bundle adjustment. We demonstrate the effectiveness of our method on KITTI and Cityscapes, where we outperform previous self-supervised approaches on multiple tasks. We also show good generalization for transfer learning.
	16	Proceedings of the IEEE ICCV, 2019, pp.7244-7253	Event-Based Motion Segmentation by Motion Compensation	Timo Stoffregen, Guillermo Gallego, Tom Drummond, Lindsay Kleeman, Davide Scaramuzza	Dept. Electrical and Computer Systems Engineering, Monash University, Australia.	不相关	In contrast to traditional cameras, whose pixels have a common exposure time, event-based cameras are novel bio-inspired sensors whose pixels work independently and asynchronously output intensity changes (called "events"), with microsecond resolution. Since events are caused by the apparent motion of objects, event-based cameras sample visual information based on the scene dynamics and are, therefore, a more natural fit than traditional cameras to acquire motion, especially at high speeds, where traditional cameras suffer from motion blur. However, distinguishing between events caused by different moving objects and by the camera's ego-motion is a challenging task. We present the first per-event segmentation method for splitting a scene into independently moving objects. Our method jointly estimates the event-object associations (i.e., segmentation) and the motion parameters of the objects (or the background) by maximization of an objective function, which builds upon recent results on event-based motion-compensation. We provide a thorough evaluation of our method on a public dataset, outperforming the state-of-the-art by as much as 10%. We also show the first quantitative evaluation of a segmentation algorithm for event cameras, yielding around 90% accuracy at 4 pixels relative displacement.
	17	Proceedings of the IEEE ICCV, 2019, pp.7525-7534	Expert Sample Consensus Applied to Camera Re-Localization	Eric Brachmann, Carsten Rother	Visual Learning Lab Heidelberg University (HCI/IWR)	不相关	Fitting model parameters to a set of noisy data points is a common problem in computer vision. In this work, we fit the 6D camera pose to a set of noisy correspondences between the 2D input image and a known 3D environment. We estimate these correspondences from the image using a neural network. Since the correspondences often contain outliers, we utilize a robust estimator such as Random Sample Consensus (RANSAC) or Differentiable RANSAC (DSAC) to fit the pose parameters. When the problem domain, e.g. the space of all 2D-3D correspondences, is large or ambiguous, a single network does not cover the domain well. Mixture of Experts (MoE) is a popular strategy to divide a problem domain among an ensemble of specialized networks, so called experts, where a gating network decides which expert is responsible for a given input. In this work, we introduce Expert Sample Consensus (ESAC), which integrates DSAC in a MoE. Our main technical contribution is an efficient method to train ESAC jointly and end-to-end. We demonstrate experimentally that ESAC handles two real-world problems better than competing methods, i.e. scalability and ambiguity. We apply ESAC to fitting simple geometric models to synthetic images, and to camera re-localization for difficult, real datasets.
	18	Proceedings of the IEEE ICCV, 2019, pp.7628-7637	Learning Single Camera Depth Estimation Using Dual-Pixels	Rahul Garg, Neal Wadhwa, Sameer Ansari, Jonathan T. Barron	Google Research	不相关	Deep learning techniques have enabled rapid progress in monocular depth estimation, but their quality is limited by the ill-posed nature of the problem and the scarcity of high quality datasets. We estimate depth from a single cam-era by leveraging the dual-pixel auto-focus hardware that is increasingly common on modern camera sensors. Classic stereo algorithms and prior learning-based depth estimation techniques underperform when applied on this dual-pixel data, the former due to too-strong assumptions about RGB image matching, and the latter due to not leveraging the understanding of optics of dual-pixel image formation. To allow learning based methods to work well on dual-pixel imagery, we identify an inherent ambiguity in the depth estimated from dual-pixel cues, and develop an approach to estimate depth up to this ambiguity. Using our approach, existing monocular depth estimation techniques can be effectively applied to dual-pixel data, and much smaller models can be constructed that still infer high quality depth. To demonstrate this, we capture a large dataset of in-the-wild 5-viewpoint RGB images paired with corresponding dual-pixel data, and show how view supervision with this data can be used to learn depth up to the unknown ambiguities. On our new task, our model is 30% more accurate than any prior work on learning-based monocular or stereoscopic depth estimation.
	19	Proceedings of the IEEE ICCV, 2019, pp.7728-7738	xR-EgoPose: Egocentric 3D Human Pose From an HMD Camera	Denis Tome, Patrick Peluse, Lourdes Agapito, Hernan Badino	University College London	相关(CNN)	We present a new solution to egocentric 3D body pose estimation from monocular images captured from a downward looking fish-eye camera installed on the rim of a head mounted virtual reality device. This unusual viewpoint, just 2 cm. away from the user's face, leads to images with unique visual appearance, characterized by severe self-occlusions and strong perspective distortions that result in a drastic difference in resolution between lower and upper body. Our contribution is two-fold. Firstly, we propose a new encoder-decoder architecture with a novel dual branch decoder designed specifically to account for the varying uncertainty in the 2D joint locations. Our quantitative evaluation, both on synthetic and real-world datasets, shows that our strategy leads to substantial improvements in accuracy over state of the art egocentric pose estimation approaches. Our second contribution is a new large-scale photorealistic synthetic dataset -- xR-EgoPose -- offering 383K frames of high quality renderings of people with a diversity of skin tones, body shapes, clothing, in a variety of backgrounds and lighting conditions, performing a range of actions. Our experiments show that the high variability in our new synthetic training corpus leads to good generalization to real world footage and to state of the art results on real world datasets with ground truth. Moreover, an evaluation on the Human3.6M benchmark shows that the performance of our method is on par with top performing approaches on the more classic problem of 3D human pose from a third person viewpoint.

	20	Proceedings of the IEEE ICCV, 2019, pp.7880-7888	Stochastic Exposure Coding for Handling Multi-ToF-Camera Interference	Jongho Lee, Mohit Gupta	University ofWisconsin-Madison	不相关	As continuous-wave time-of-flight (C-ToF) cameras become popular in 3D imaging applications, they need to contend with the problem of multi-camera interference (MCI). In a multi-camera environment, a ToF camera may receive light from the sources of other cameras, resulting in large depth errors. In this paper, we propose stochastic exposure coding (SEC), a novel approach for mitigating. SEC involves dividing a camera's integration time into multiple slots, and switching the camera off and on stochastically during each slot. This approach has two benefits. First, by appropriately choosing the on probability for each slot, the camera can effectively filter out both the AC and DC components of interfering signals, thereby mitigating depth errors while also maintaining high signal-to-noise ratio. This enables high accuracy depth recovery with low power consumption. Second, this approach can be implemented without modifying the C-ToF camera's coding functions, and thus, can be used with a wide range of cameras with minimal changes. We demonstrate the performance benefits of SEC with theoretical analysis, simulations and real experiments, across a wide range of imaging scenarios.
	21	Proceedings of the IEEE ICCV, 2019, pp.8080-8089	A Novel Unsupervised Camera-Aware Domain Adaptation Framework for Person Re-Identification	Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, Yang Gao	State Key Laboratory for Novel Software Technology, Nanjing University	不相关	Unsupervised cross-domain person re-identification (Re-ID) faces two key issues. One is the data distribution discrepancy between source and target domains, and the other is the lack of discriminative information in target domain. From the perspective of representation learning, this paper proposes a novel end-to-end deep domain adaptation framework to address them. For the first issue, we highlight the presence of camera-level sub-domains as a unique characteristic in person Re-ID, and develop a "camera-aware" domain adaptation method via adversarial learning. With this method, the learned representation reduces distribution discrepancy not only between source and target domains but also across all cameras. For the second issue, we exploit the temporal continuity in each camera of target domain to create discriminative information. This is implemented by dynamically generating online triplets within each batch, in order to maximally take advantage of the steadily improved representation in training process. Together, the above two methods give rise to a new unsupervised domain adaptation framework for person Re-ID. Extensive experiments and ablation studies conducted on benchmark datasets demonstrate its superiority and interesting properties.
	22	Proceedings of the IEEE ICCV, 2019, pp.8908-8917	Watch, Listen and Tell: Multi-Modal Weakly Supervised Dense Event Captioning	Tanzila Rahman, Bicheng Xu, Leonid Sigal	University of British Columbia	不相关	Multi-modal learning, particularly among imaging and linguistic modalities, has made amazing strides in many high-level fundamental visual understanding problems, ranging from language grounding to dense event captioning. However, much of the research has been limited to approaches that either do not take audio corresponding to video into account at all, or those that model the audio-visual correlations in service of sound or sound source localization. In this paper, we present the evidence, that audio signals can carry surprising amount of information when it comes to high-level visual-lingual tasks. Specifically, we focus on the problem of weakly-supervised dense event captioning in videos and show that audio on its own can nearly rival performance of a state-of-the-art visual model and, combined with video, can improve on the state-of-the-art performance. Extensive experiments on the ActivityNet Captions dataset show that our proposed multi-modal approach outperforms state-of-the-art unimodal methods, as well as validate
	23	Proceedings of the IEEE ICCV, 2019, pp.9308-9318	WoodScape: A Multi-Task, Multi-Camera Fisheye Dataset for Autonomous Driving	Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sumanth Chennupati, Sanjaya Nayak, Saquib Mansoor, Xavier	/	相关(CNN)	Fisheye cameras are commonly employed for obtaining a large field of view in surveillance, augmented reality and in particular automotive applications. In spite of their prevalence, there are few public datasets for detailed evaluation of computer vision algorithms on fisheye images. We release the first extensive fisheye automotive dataset, WoodScape, named after Robert Wood who invented the fisheye camera in 1906. WoodScape comprises of four surround view cameras and nine tasks including segmentation, depth estimation, 3D bounding box detection and soiling detection. Semantic annotation of 40 classes at the instance level is provided for over 10,000 images and annotation for other tasks are provided for over 100,000 images. With WoodScape, we would like to encourage the community to adapt computer vision models for fisheye camera instead of using naive rectification.
	24	Proceedings of the IEEE ICCV, 2019, pp.9974-9983	UprightNet: Geometry-Aware Camera Orientation Estimation From Single Images	Wenqi Xian, Zhengqi Li, Matthew Fisher, Jonathan Eisenmann, Eli Shechtman, Noah Snavely	Cornell Tech, Cornell University	不相关	We introduce UprightNet, a learning-based approach for estimating 2DoF camera orientation from a single RGB image of an indoor scene. Unlike recent methods that leverage deep learning to perform black-box regression from image to orientation parameters, we propose an end-to-end framework that incorporates explicit geometric reasoning. In particular, we design a network that predicts two representations of scene geometry, in both the local camera and global reference coordinate systems, and solves for the camera orientation as the rotation that best aligns these two predictions via a differentiable least squares module. This network can be trained end-to-end, and can be supervised with both ground truth camera poses and intermediate representations of surface geometry. We evaluate UprightNet on the single-image camera orientation task on synthetic and real datasets, and show significant improvements over prior state-of-the-art approaches.
	25	Proceedings of the IEEE ICCV, 2019, pp.10133-10142	Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image	Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee	ECE & ASRI, Seoul National University, Korea	不相关	Although significant improvement has been achieved recently in 3D human pose estimation, most of the previous methods only treat a single-person case. In this work, we firstly propose a fully learning-based, camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. The pipeline of the proposed system consists of human detection, absolute 3D human root localization, and root-relative 3D single-person pose estimation modules. Our system achieves comparable results with the state-of-the-art 3D single-person pose estimation models without any groundtruth information and significantly outperforms previous 3D multi-person pose estimation methods on publicly available datasets. The code is available in (https://github.com/mks0601/3DMPPE_ROOTNET_RELEASE) , (https://github.com/mks0601/3DMPPE_POSENET_RELEASE).
	1	Proceedings of the IEEE ICCV, 2021, pp.12385-12395	EventHands: Real-Time Neural 3D Hand Pose Estimation From an Event Stream	Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, Christian Theobalt	MPI for Informatics, SIC	不相关	3D hand pose estimation from monocular videos is a long-standing and challenging problem, which is now seeing a strong upturn. In this work, we address it for the first time using a single event camera, i.e., an asynchronous vision sensor reacting on brightness changes. Our EventHands approach has characteristics previously not demonstrated with a single RGB or depth camera such as high temporal resolution at low data throughputs and real-time performance at 1000 Hz. Due to the different data modality of event cameras compared to classical cameras, existing methods cannot be directly applied to and re-trained for event streams. We thus design a new neural approach which accepts a new event stream representation suitable for learning, which is trained on newly-generated synthetic event streams and can generalise to real data. Experiments show that EventHands outperforms recent monocular methods using a colour (or depth) camera in terms of accuracy and its ability to capture hand motions of unprecedented speed. Our method, the event stream simulator and the dataset are publicly available (see https://gvv.mpi-inf.mpg.de/projects/EventHands/).

2	Proceedings of the IEEE ICCV, 2021, pp.12558-12567	In-the-Wild Single Camera 3D Reconstruction Through Moving Water Surfaces	Jinhui Xiong, Wolfgang Heidrich	KAUST	不相关	We present a method for reconstructing the 3D shape of underwater environments from a single, stationary camera placed above the water. We propose a novel differentiable framework, which, to our knowledge, is the first single-camera solution that is capable of simultaneously retrieving the structure of dynamic water surfaces and static underwater scene geometry in the wild. This framework integrates ray casting of Snell's law at the refractive interface, multi-view triangulation and specially designed loss functions. Our method is calibration-free, and thus it is easy to collect data outdoors in uncontrolled environments. Experimental results show that our method is able to realize robust and quality reconstructions on a variety of scenes, both in a laboratory environment and in the wild, and even in a salt water environment. We believe the method is promising for applications
3	Proceedings of the IEEE ICCV, 2021, pp.448-457	An Asynchronous Kalman Filter for Hybrid Event Cameras	Ziwei Wang, Yonhon Ng, Cedric Scheerlinck, Robert Mahony	The Australian National University Systems Theory and Robotics Group	不相关	Event cameras are ideally suited to capture HDR visual information without blur but perform poorly on static or slowly changing scenes. Conversely, conventional image sensors measure absolute intensity of slowly changing scenes effectively but do poorly on high dynamic range or quickly changing scenes. In this paper, we present an event-based video reconstruction pipeline for High Dynamic Range (HDR) scenarios. The proposed algorithm includes a frame augmentation pre-processing step that deblurs and temporally interpolates frame data using events. The augmented frame and event data are then fused using a novel asynchronous Kalman filter under a unifying uncertainty model for both sensors. Our experimental results are evaluated on both publicly available datasets with challenging lighting conditions and fast motions and our new dataset with HDR reference. The proposed algorithm outperforms state-of-the-art methods in both absolute intensity error (48% reduction) and image similarity indexes (average 11% improvement).
4	Proceedings of the IEEE ICCV, 2021, pp.4955-4964	The Benefit of Distraction: Denoising Camera-Based Physiological Measurements Using Inverse Attention	Ewa M. Nowara, Daniel McDuff, Ashok Veeraraghavan	Rice University, Houston, TX	不相关	Attention networks perform well on diverse computer vision tasks. The core idea is that the signal of interest is stronger in some pixels ("foreground"), and by selectively focusing computation on these pixels, networks can extract subtle information buried in noise and other sources of corruption. Our paper is based on one key observation: in many real-world applications, many sources of corruption, such as illumination and motion, are often shared between the "foreground" and the "background" pixels. Can we utilize this to our advantage? We propose the utility of inverse attention networks, which focus on extracting information about these shared sources of corruption. We show that this helps to effectively suppress shared covariates and amplify signal information, resulting in improved performance. We illustrate this on the task of camera-based physiological measurement where the signal of interest is weak and global illumination variations and motion act as significant shared sources of corruption. We perform experiments on three datasets and show that our approach of inverse attention produces state-of-the-art results, increasing the signal-to-noise ratio by up to 5.8 dB, reducing heart rate and breathing rate estimation errors by as much as 30 %, recovering subtle waveform dynamics, and generalizing from RGB to NIR videos without retraining.
5	Proceedings of the IEEE ICCV, 2021, pp.4480-4489	Event Stream Super-Resolution via Spatiotemporal Constraint Learning	Siqi Li, Yutong Feng, Yipeng Li, Yu Jiang, Changqing Zou, Yue Gao	BNRist, THUICBS, KLISS, School of Software, Tsinghua University, China	相关(SNN)	Event cameras are bio-inspired sensors that respond to brightness changes asynchronously and output in the form of event streams instead of frame-based images. They own outstanding advantages compared with traditional cameras: higher temporal resolution, higher dynamic range, and lower power consumption. However, the spatial resolution of existing event cameras is insufficient and challenging to be enhanced at the hardware level while maintaining the asynchronous philosophy of circuit design. Therefore, it is imperative to explore the algorithm of event stream super-resolution, which is a non-trivial task due to the sparsity and strong spatio-temporal correlation of the events from an event camera. In this paper, we propose an end-to-end framework based on spiking neural network for event stream super-resolution, which can generate high-resolution (HR) event stream from the input low-resolution (LR) event stream. A spatiotemporal constraint learning mechanism is proposed to learn the spatial and temporal distributions of the event stream simultaneously. We validate our method on four large-scale datasets and the results show that our method achieves state-of-the-art performance. The satisfying results on two downstream applications, i.e. object classification and image reconstruction, further demonstrate the usability of our method. To prove the application potential of our method, we deploy it on a mobile platform. The high-quality HR event stream generated by our real-time system demonstrates the effectiveness and efficiency of our method.
6	Proceedings of the IEEE ICCV, 2021, pp.3252-3262	Continual Learning for Image-Based Camera Localization	Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, Juho Kannala	Aalto University	不相关	For several emerging technologies such as augmented reality, autonomous driving and robotics, visual localization is a critical component. Directly regressing camera pose/3D scene coordinates from the input image using deep neural networks has shown great potential. However, such methods assume a stationary data distribution with all scenes simultaneously available during training. In this paper, we approach the problem of visual localization in a continual learning setup -- whereby the model is trained on scenes in an incremental manner. Our results show that similar to the classification domain, non-stationary data induces catastrophic forgetting in deep networks for visual localization. To address this issue, a strong baseline based on storing and replaying images from a fixed buffer is proposed. Furthermore, we propose a new sampling method based on coverage score (Buff-CS) that adapts the existing sampling strategies in the buffering process to the problem of visual localization. Results demonstrate consistent improvements over standard buffering methods on two challenging datasets -- 7Scenes, 12Scenes, and also 19Scenes by combining the former scenes.
7	Proceedings of the IEEE ICCV, 2021, pp.2430-2439	A Dark Flash Normal Camera	Zhihao Xia, Jason Lawrence, Supreeth Achar	Washington University in St. Louis	不相关	Casual photography is often performed in uncontrolled lighting that can result in low quality images and degrade the performance of downstream processing. We consider the problem of estimating surface normal and reflectance maps of scenes depicting people despite these conditions by supplementing the available visible illumination with a single near infrared (NIR) light source and camera, a so-called "dark flash image". Our method takes as input a single color image captured under arbitrary visible lighting and a single dark flash image captured under controlled front-lit NIR lighting at the same viewpoint, and computes a normal map, a diffuse albedo map, and a specular intensity map of the scene. Since ground truth normal and reflectance maps of faces are difficult to capture, we propose a novel training technique that combines information from two readily available and complementary sources: a stereo depth signal and photometric shading cues. We evaluate our method over a range of subjects and lighting conditions and describe two applications: optimizing stereo geometry and filling the shadows in an image.
8	Proceedings of the IEEE ICCV, 2021, pp.934-943	Graph-Based Asynchronous Event Processing for Rapid Object Recognition	Yijin Li, Han Zhou, Bangbang Yang, Ye Zhang, Zhaopeng Cui, Hujun Bao, Guofeng Zhang	State Key Lab of CAD&CG, Zhejiang University	不相关	Different from traditional video cameras, event cameras capture asynchronous events stream in which each event encodes pixel location, trigger time, and the polarity of the brightness changes. In this paper, we introduce a novel graph-based framework for event cameras, namely SlideGCN. Unlike some recent graph-based methods that use groups of events as input, our approach can efficiently process data event-by-event, unlock the low latency nature of events data while still maintaining the graph's structure internally. For fast graph construction, we develop a radius search algorithm, which better exploits the partial regular structure of event cloud against k-d tree based generic methods. Experiments show that our method reduces the computational complexity up to 100 times with respect to current graph-based methods while keeping state-of-the-art performance on object recognition. Moreover, we verify the superiority of event-wise processing with our method. When the state becomes stable, we can give a prediction with high confidence, thus

9	Proceedings of the IEEE ICCV, 2021, pp.15263-15272	Robust Small Object Detection on the Water Surface Through Fusion of Camera and Millimeter Wave Radar	Yuwei Cheng, Hu Xu, Yimin Liu	Tsinghua University	相关(LiDAR)	In recent years, unmanned surface vehicles (USVs) have been experiencing growth in various applications. With the expansion of USVs' application scenes from the typical marine areas to inland waters, new challenges arise for the object detection task, which is an essential part of the perception system of USVs. In our work, we focus on a relatively unexplored task for USVs in inland waters: small object detection on water surfaces, which is of vital importance for safe autonomous navigation and USVs' certain missions such as floating waste cleaning. Considering the limitations of vision-based object detection, we propose a novel vision-radar fusion based method for robust small object detection on water surfaces. By using a novel representation format of millimeter wave radar point clouds and applying a deep-level multi-scale fusion of RGB images and radar data, the proposed method can efficiently utilize the characteristics of radar data and improve the accuracy and robustness for small object detection on water surfaces. We test the method on the real-world floating bottle dataset that we collected and released. The result shows that, our method improves the average detection accuracy significantly compared to the vision-based methods and achieves state-of-the-art performance. Besides, the proposed method performs robustly when single sensor degrades.
10	Proceedings of the IEEE ICCV, 2021, pp.6351-6361	GNeRF: GAN-Based Neural Radiance Field Without Posed Camera	Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, Jingyi Yu	Shanghai Engineering Research Center of Intelligent Vision and Imaging	不相关	We introduce GNeRF, a framework to marry Generative Adversarial Networks (GAN) with Neural Radiance Field (NeRF) reconstruction for the complex scenarios with unknown and even randomly initialized camera poses. Recent NeRF-based advances have gained popularity for remarkable realistic novel view synthesis. However, most of them heavily rely on accurate camera poses estimation, while few recent methods can only optimize the unknown camera poses in roughly forward-facing scenes with relatively short camera trajectories and require rough camera poses initialization. Differently, our GNeRF only utilizes randomly initialized poses for complex outside-in scenarios. We propose a novel two-phases end-to-end framework. The first phase takes the use of GANs into the new realm for optimizing coarse camera poses and radiance fields jointly, while the second phase refines them with additional photometric loss. We overcome local minima using a hybrid and iterative optimization scheme. Extensive experiments on a variety of synthetic and natural scenes demonstrate the effectiveness of GNeRF. More impressively, our approach outperforms the baselines favorably in those scenes with repeated patterns or even low textures that are regarded as extremely challenging before.
11	Proceedings of the IEEE ICCV, 2021, pp.13495-13504	The Spatio-Temporal Poisson Point Process: A Simple Model for the Alignment of Event Camera Data	Cheng Gu, Erik Learned-Miller, Daniel Sheldon, Guillermo Gallego, Pia Bideau	TU Berlin	不相关	Event cameras, inspired by biological vision systems, provide a natural and data efficient representation of visual information. Visual information is acquired in the form of events that are triggered by local brightness changes. However, because most brightness changes are triggered by relative motion of the camera and the scene, the events recorded at a single sensor location seldom correspond to the same world point. To extract meaningful information from event cameras, it is helpful to register events that were triggered by the same underlying world point. In this work we propose a new model of event data that captures its natural spatio-temporal structure. We start by developing a model for aligned event data. That is, we develop a model for the data as though it has been perfectly registered already. In particular, we model the aligned data as a spatio-temporal Poisson point process. Based on this model, we develop a maximum likelihood approach to registering events that are not yet aligned. That is, we find transformations of the observed events that make them as likely as possible under our model. In particular we extract the camera rotation that leads to the best event alignment. We show new state of the art accuracy for rotational velocity estimation on the DAVIS 240C dataset [??]. In addition, our method is also faster and has lower computational complexity than several competing methods.
12	Proceedings of the IEEE ICCV, 2021, pp.4882-4891	EvIntSR-Net: Event Guided Multiple Latent Frames Reconstruction and Super-Resolution	Jin Han, Yixin Yang, Chu Zhou, Chao Xu, Boxin Shi	Key Lab of Machine Perception (MOE), Dept. of Machine Intelligence, Peking University	不相关	An event camera detects the scene radiance changes and sends a sequence of asynchronous event streams with high dynamic range, high temporal resolution, and low latency. However, the spatial resolution of event cameras is limited as a trade-off for these outstanding properties. To reconstruct high-resolution intensity images from event data, we propose EvIntSR-Net that converts event data to multiple latent intensity frames to achieve super-resolution on intensity images in this paper. EvIntSR-Net bridges the domain gap between event streams and intensity frames and learns to merge a sequence of latent intensity frames in a recurrent updating manner. Experimental results show that EvIntSR-Net can reconstruct SR intensity images with higher dynamic range and fewer blurry artifacts by fusing events with intensity frames for both simulated and real-world data. Furthermore, the proposed EvIntSR-Net is able to generate high-frame-rate videos with super-resolved frames.
13	Proceedings of the IEEE ICCV, 2021, pp.4248-4257	ReconfigISP: Reconfigurable Camera Image Processing Pipeline	Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, Jinwei Gu	SenseTime Research and Tetras.AI	不相关	Image Signal Processor (ISP) is a crucial component in digital cameras that transforms sensor signals into images for us to perceive and understand. Existing ISP designs always adopt a fixed architecture, e.g., several sequential modules connected in a rigid order. Such a fixed ISP architecture may be suboptimal for real-world applications, where camera sensors, scenes and tasks are diverse. In this study, we propose a novel Reconfigurable ISP (ReconfigISP) whose architecture and parameters can be automatically tailored to specific data and tasks. In particular, we implement several ISP modules, and enable backpropagation for each module by training a differentiable proxy, hence allowing us to leverage the popular differentiable neural architecture search and effectively search for the optimal ISP architecture. A proxy tuning mechanism is adopted to maintain the accuracy of proxy networks in all cases. Extensive experiments conducted on image restoration and object detection, with different sensors, light conditions and efficiency constraints, validate the effectiveness of ReconfigISP. Only hundreds of parameters need tuning for every task.
14	Proceedings of the IEEE ICCV, 2021, pp.2563-2572	Event-Based Video Reconstruction Using Transformer	Wenming Weng, Yueyi Zhang, Zhiwei Xiong	University of Science and Technology of China	相关(CNN)	Event cameras, which output events by detecting spatio-temporal brightness changes, bring a novel paradigm to image sensors with high dynamic range and low latency. Previous works have achieved impressive performances on event-based video reconstruction by introducing convolutional neural networks (CNNs). However, intrinsic locality of convolutional operations is not capable of modeling long-range dependency, which is crucial to many vision tasks. In this paper, we present a hybrid CNN-Transformer network for event-based video reconstruction (ET-Net), which merits the fine local information from CNN and global contexts from Transformer. In addition, we further propose a Token Pyramid Aggregation strategy to implement multi-scale token integration for relating internal and intersected semantic concepts in the token-space. Experimental results demonstrate that our proposed method achieves superior performance over state-of-the-art methods on multiple real-world event datasets. The code is available at

2021	15	Proceedings of the IEEE ICCV, 2021, pp.6148-6157	Rational Polynomial Camera Model Warping for Deep Learning Based Satellite Multi-View Stereo Matching	Jian Gao, Jin Liu, Shunping Ji	School of Remote Sensing and information Engineering, Wuhan University, China	不相关	Satellite multi-view stereo (MVS) imagery is particularly suited for large-scale Earth surface reconstruction. Differing from the perspective camera model (pin-hole model) that is commonly used for close-range and aerial cameras, the cubic rational polynomial camera (RPC) model is the mainstream model for push-broom linear-array satellite cameras. However, the homography warping used in the prevailing learning based MVS methods is only applicable to pin-hole cameras. In order to apply the SOTA learning based MVS technology to the satellite MVS taskfor large-scale Earth surface reconstruction, RPC warping should be considered. In this work, we propose, for the first time, a rigorous RPC warping module. The rational polynomial coefficients are recorded as a tensor, and the RPC warping is formulated as a series of tensor transformations. Based on the RPC warping, we propose the deep learning based satellite MVS (SatMVS) framework for large-scale and wide depth range Earth surface reconstruction. We also introduce a large-scale satellite image dataset consisting of 519 5120x5120 images, which we call the TLC SatMVS dataset. The satellite images were acquired from a three-line camera (TLC) that catches triple-view images simultaneously, forming a valuable supplement to the existing open-source WorldView-3 datasets with single-scanline images. Experiments show that the proposed RPC warping module and the SatMVS framework can achieve a superior reconstruction accuracy compared to the pin-hole fitting method and conventional MVS methods. Code and data are available at https://github.com/WHU-GPCV/SatMVS.
	16	Proceedings of the IEEE ICCV, 2021, pp.13043-13052	Object Tracking by Jointly Exploiting Frame and Event Domain	Jiqing Zhang, Xin Yang, Yingkai Fu, Xiaopeng Wei, Baocai Yin, Bo Dong	Dalian University of Technology	相关(CNN)	Inspired by the complementarity between conventional frame-based and bio-inspired event-based cameras, we propose a multi-modal based approach to fuse visual cues from the frame- and event-domain to enhance the single object tracking performance, especially in degraded conditions (e.g., scenes with high dynamic range, low light, and fast-motion objects). The proposed approach can effectively and adaptively combine meaningful information from both domains. Our approach's effectiveness is enforced by a novel designed cross-domain attention schemes, which can effectively enhance features based on self- and cross-domain attention schemes; The adaptiveness is guarded by a specially designed weighting scheme, which can adaptively balance the contribution of the two domains. To exploit event-based visual cues in single-object tracking, we construct a large-scale frame-event-based dataset, which we subsequently employ to train a novel frame-event fusion based model. Extensive experiments show that the proposed approach outperforms state-of-the-art frame-based tracking methods by at least 10.4% and 11.9% in terms of representative success rate and precision rate, respectively. Besides, the effectiveness of each key component of our approach is evidenced by our thorough ablation study.
	17	Proceedings of the IEEE ICCV, 2021, pp.6218-6228	On the Limits of Pseudo Ground Truth in Visual Camera Re-Localisation	Eric Brachmann, Martin Humenberger, Carsten Rother, Torsten Sattler	Niantic	不相关	Benchmark datasets that measure camera pose accuracy have driven progress in visual re-localisation research. To obtain poses for thousands of images, it is common to use a reference algorithm to generate pseudo ground truth. Popular choices include Structure-from-Motion (SfM) and Simultaneous-Localisation-and-Mapping (SLAM) using additional sensors like depth cameras if available. Re-localisation benchmarks thus measure how well each method replicates the results of the reference algorithm. This begs the question whether the choice of the reference algorithm favours a certain family of re-localisation methods. This paper analyzes two widely used re-localisation datasets and shows that evaluation outcomes indeed vary with the choice of the reference algorithm. We thus question common beliefs in the re-localisation literature, namely that learning-based scene coordinate regression outperforms classical feature-based methods, and that RGB-D- based methods outperform RGB-based methods. We argue that any claims on ranking re-localisation methods should take the type of the reference algorithm, and the similarity of the methods to the reference algorithm, into account.
	18	Proceedings of the IEEE ICCV, 2021, pp.4228-4237	Inverting a Rolling Shutter Camera: Bring Rolling Shutter Images to High Framerate Global Shutter Video	Bin Fan, Yuchao Dai	School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China	不相关	Rolling shutter (RS) images can be viewed as the result of the row-wise combination of global shutter (GS) images captured by a virtual moving GS camera over the period of camera readout time. The RS effect brings tremendous difficulties for the downstream applications. In this paper, we propose to invert the above RS imaging mechanism, i.e., recovering a high framerate GS video from consecutive RS images to achieve RS temporal super-resolution (RSSR). This extremely challenging problem, e.g., recovering 1440 GS images from two 720-height RS images, is far from being solved end-to-end. To address this challenge, we exploit the geometric constraint in the RS camera model, thus achieving geometry-aware inversion. Specifically, we make three contributions in resolving the above difficulties: (i) formulating the bidirectional RS undistortion flows under the constant velocity motion model, (ii) building the connection between the RS undistortion flow and optical flow via a scaling operation, and (iii) developing a mutual conversion scheme between varying RS undistortion flows that correspond to different scanlines. Building upon these formulations, we propose the first RS temporal super-resolution network in a cascaded structure to extract high framerate global shutter video. Our method explores the underlying spatio-temporal geometric relationships within a deep learning framework, where no extra supervision besides the middle-scanline ground truth GS image is needed. Essentially, our method can be very efficient for explicit propagation to generate GS images under any scanline. Experimental results on both synthetic and real data show that our method can produce high-quality GS image sequences with rich details,
	19	Proceedings of the IEEE ICCV, 2021, pp.1981-1990	Cross-Camera Convolutional Color Constancy	Mahmoud Afifi, Jonathan T. Barron, Chloe LeGendre, Yun-Ta Tsai, Francois Bleibel	Google Research	不相关	We present "Cross-Camera Convolutional Color Constancy" (C5), a learning-based method, trained on images from multiple cameras, that accurately estimates a scene's illuminant color from raw images captured by a new camera previously unseen during training. C5 is a hypernetwork-like extension of the convolutional color constancy (CCC) approach: C5 learns to generate the weights of a CCC model that is then evaluated on the input image, with the CCC weights dynamically adapted to different input content. Unlike prior cross-camera color constancy models, which are usually designed to be agnostic to the spectral properties of test-set images from unobserved cameras, C5 approaches this problem through the lens of transductive inference: additional unlabeled images are provided as input to the model at test time, which allows the model to calibrate itself to the spectral properties of the test-set camera during inference. C5 achieves state-of-the-art accuracy for cross-camera color constancy on several datasets, is fast to evaluate (7 and 90 ms per image on a GPU or CPU, respectively), and requires little memory (2 MB), and thus is a practical solution to the problem of calibration-free automatic white balance for
	20	Proceedings of the IEEE ICCV, 2021, pp.2146-2156	N-ImageNet: Towards Robust, Fine-Grained Object Recognition With Event Cameras	Junho Kim, Jaehyeok Bae, Gangin Park, Dongsu Zhang, Young Min Kim	Dept. of Electrical and Computer Engineering, Seoul National University, Korea	不相关	We introduce N-ImageNet, a large-scale dataset targeted for robust, fine-grained object recognition with event cameras. The dataset is collected using programmable hardware in which an event camera consistently moves around a monitor displaying images from ImageNet. N-ImageNet serves as a challenging benchmark for event-based object recognition, due to its large number of classes and samples. We empirically show that pretraining on N-ImageNet improves the performance of event-based classifiers and helps them learn with few labeled data. In addition, we present several variants of N-ImageNet to test the robustness of event-based classifiers under diverse camera trajectories and severe lighting conditions, and propose a novel event representation to alleviate the performance degradation. To the best of our knowledge, we are the first to quantitatively investigate the consequences caused by various environmental conditions on event-based object recognition algorithms. N-ImageNet and its variants are expected to guide practical implementations for deploying event-based object recognition algorithms in the real world.

21	Proceedings of the IEEE ICCV, 2021, pp.4258-4267	Event-Intensity Stereo: Estimating Depth by the Best of Both Worlds	Mohammad Mostafavi, Kuk-Jin Yoon, Jonghyun Choi	GIST, South Korea	不相关	Event cameras can report scene movements as an asynchronous stream of data called the events. Unlike traditional cameras, event cameras have very low latency (microseconds vs milliseconds) very high dynamic range (140dB vs 60 dB), and low power consumption, as they report changes of a scene and not a complete frame. As they re-port per pixel feature-like events and not the whole intensity frame they are immune to motion blur. However, event cameras require movement between the scene and camera to fire events ,i.e., they have no output when the scene is relatively static. Traditional cameras, however, report the whole frame of pixels at once in fixed intervals but have lower dynamic range and are prone to motion blur in case of rapid movements. We get the best from both worlds and use events and intensity images together in our complementary design and estimate dense disparity from this combination. The proposed end-to-end design combines events and images in a sequential manner and correlates them to esti-mate dense depth values. Our various experimental settings in real-world and simulated scenarios exploit the superiority of our method in predicting accurate depth values with fine details. We further extend our method to extreme cases of missing the left or right event or stereo pair and also investigate stereo depth estimation with inconsistent dynamic ranges or event
22	Proceedings of the IEEE ICCV, 2021, pp.11169-11178	Camera Distortion-Aware 3D Human Pose Estimation in Video With Optimization-Based Meta-Learning	Hanbyel Cho, Yooshin Cho, Jaemyung Yu, Junmo Kim	School of Electrical Engineering, KAIST, South Korea	不相关	Existing 3D human pose estimation algorithms trained on distortion-free datasets suffer performance drop when applied to new scenarios with a specific camera distortion. In this paper, we propose a simple yet effective model for 3D human pose estimation in video that can quickly adapt to any distortion environment by utilizing MAML, a representative optimization-based meta-learning algorithm. We consider a sequence of 2D keypoints in a particular distortion as a single task of MAML. However, due to the absence of a large-scale dataset in a distorted environment, we propose an efficient method to generate synthetic distorted data from undistorted 2D keypoints. For the evaluation, we assume two practical testing situations depending on whether a motion capture sensor is available or not. In particular, we propose Inference Stage Optimization using bone-length symmetry and consistency. Extensive evaluation shows that our proposed method successfully adapts to various degrees of distortion in the testing phase and outperforms the existing state-of-the-art approaches. The proposed method is useful in practice because it does not require camera calibration and additional computations in a testing set-up.
23	Proceedings of the IEEE ICCV, 2021, pp.2001-2010	Dual-Camera Super-Resolution With Aligned Attention Modules	Tengfei Wang, Jiaxin Xie, Wenxiu Sun, Qiong Yan, Qifeng Chen	HKUST	不相关	We present a novel approach to reference-based super-resolution (RefSR) with the focus on dual-camera super-resolution (DCSR), which utilizes reference images for high-quality and high-fidelity results. Our proposed method generalizes the standard patch-based feature matching with spatial alignment operations. We further explore the dual-camera super-resolution that is one promising application of RefSR, and build a dataset that consists of 146 image pairs from the main and telephoto cameras in a smartphone. To bridge the domain gaps between real-world images and the training images, we propose a self-supervised domain adaptation strategy for real-world images. Extensive experiments on our dataset and a public benchmark demonstrate clear improvement achieved by our method over state of the art in both quantitative evaluation and visual comparisons.
24	Proceedings of the IEEE ICCV, 2021, pp.15631-15640	Robust 2D/3D Vehicle Parsing in Arbitrary Camera Views for CVIS	Hui Miao, Feixiang Lu, Zongdai Liu, Liangjun Zhang, Dinesh Manocha, Bin Zhou	State Key Laboratory of Virtual Reality Technology and Systems, Beihang University	不相关	We present a novel approach to robustly detect and perceive vehicles in different camera views as part of a cooperative vehicle-infrastructure system (CVIS). Our formulation is designed for arbitrary camera views and makes no assumptions about intrinsic or extrinsic parameters. First, to deal with multi-view data scarcity, we propose a part-assisted novel view synthesis algorithm for data augmentation. We train a part-based texture inpainting network in a self-supervised manner. Then we render the textured model into the background image with the target 6-DoF pose. Second, to handle various camera parameters, we present a new method that produces dense mappings between image pixels and 3D points to perform robust 2D/3D vehicle parsing. Third, we build the first CVIS dataset for benchmarking, which annotates more than 1540 images (14017 instances) from real-world traffic scenarios. We combine these novel algorithms and datasets to develop a robust approach for 2D/3D vehicle parsing for CVIS. In practice, our approach outperforms SOTA methods on 2D detection, instance segmentation, and 6-DoF pose estimation by 3.8%, 4.3%, and 2.9%, respectively.
25	Proceedings of the IEEE ICCV, 2021, pp.9834-9844	Visio-Temporal Attention for Multi-Camera Multi-Target Association	Yu-Jhe Li, Xinshuo Weng, Yan Xu, Kris M. Kitani	Carnegie Mellon University	不相关	We address the task of Re-Identification (Re-ID) in multi-target multi-camera (MTMC) tracking where we track multiple pedestrians using multiple overlapping uncalibrated (unknown pose) cameras. Since the videos are temporally synchronized and spatially overlapping, we can see a person from multiple views and associate their trajectory across cameras. In order to find the correct association between pedestrians visible from multiple views during the same time window, we extract a visual feature from a tracklet (sequence of pedestrian images) that encodes its similarity and dissimilarity to all other candidate tracklets. We propose a inter-tracklet (person to person) attention mechanism that learns a representation for a target tracklet while taking into account other tracklets across multiple views. Furthermore, to encode the gait and motion of a person, we introduce second intra-tracklet (person-specific) attention module with position embeddings. This second module employs a transformer encoder to learn a feature from a sequence of features over one tracklet. Experimental results on WILDTRACK and our new dataset `ConstructSite' confirm the superiority of our model over state-of-the-art ReID methods (5% and 10% performance gain respectively) in the context of uncalibrated MTMC tracking. While our model is designed for overlapping cameras, we also obtain state-of-the-art results on two other benchmark datasets (MARS and DukeMTMC) with non-overlapping cameras.
26	Proceedings of the IEEE ICCV, 2021, pp.8075-8084	Generic Event Boundary Detection: A Benchmark for Event Segmentation	Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, Matt Feiszli	Facebook AI	不相关	This paper presents a novel task together with a new benchmark for detecting generic, taxonomy-free event boundaries that segment a whole video into chunks. Conventional work in temporal video segmentation and action detection focuses on localizing pre-defined action categories and thus does not scale to generic videos. Cognitive Science has known since last century that humans consistently segment videos into meaningful temporal chunks. This segmentation happens naturally, without pre-defined event categories and without being explicitly asked to do so. Here, we repeat these cognitive experiments on mainstream CV datasets; with our novel annotation guideline which addresses the complexities of taxonomy-free event boundary annotation, we introduce the task of Generic Event Boundary Detection (GEBD) and the new benchmark Kinetics-GEBD. We view GEBD as an important stepping stone towards understanding the video as a whole, and believe it has been previously neglected due to a lack of proper task definition and annotations. Through experiment and human study we demonstrate the value of the annotations. Further, we benchmark supervised and un-supervised GEBD approaches on the TAPOS dataset and our Kinetics-GEBD. We release our annotations and baseline codes at CVPR'21 LOVEU Challenge: https://sites.google.com/view/loveucvpr21.

27	Proceedings of the IEEE ICCV, 2021, pp.11035-11045	SPEC: Seeing People in the Wild With an Estimated Camera	Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, Michael J. Black	Max Planck Institute for Intelligent Systems, T¨ubingen, Germany	不相关	Due to the lack of camera parameter information for in-the-wild images, existing 3D human pose and shape (HPS) estimation methods make several simplifying assumptions: weak-perspective projection, large constant focal length, and zero camera rotation. These assumptions often do not hold and we show, quantitatively and qualitatively, that they cause errors in the reconstructed 3D shape and pose. To address this, we introduce SPEC, the first in-the-wild 3D HPS method that estimates the perspective camera from a single image and employs this to reconstruct 3D human bodies more accurately. First, we train a neural network to estimate the field of view, camera pitch, and roll given an input image. We employ novel losses that improve the calibration accuracy over previous work. We then train a novel network that concatenates the camera calibration to the image features and uses these together to regress 3D body shape and pose. SPEC is more accurate than the prior art on the standard benchmark (3DPW) as well as two new datasets with more challenging camera views and varying focal lengths. Specifically, we create a new photorealistic synthetic dataset (SPEC-SYN) with ground truth 3D bodies and a novel in-the-wild dataset (SPEC-MTP) with calibration and high-quality reference bodies. Code and datasets are available for research purposes at https://spec.is.tue.mpg.de/.
28	Proceedings of the IEEE ICCV, 2021, pp.16198-16207	Full-Velocity Radar Returns by Radar-Camera Fusion	Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, Praveen Narayanan	Michigan State University	相关(LiDAR)	A distinctive feature of Doppler radar is the measurement of velocity in the radial direction for radar points. However, the missing tangential velocity component hampers object velocity estimation as well as temporal integration of radar sweeps in dynamic scenes. Recognizing that fusing camera with radar provides complementary information to radar, in this paper we present a closed-form solution for the point-wise, full-velocity estimate of Doppler returns using the corresponding optical flow from camera images. Additionally, we address the association problem between radar returns and camera images with a neural network that is trained to estimate radar-camera correspondences. Experimental results on the nuScenes dataset verify the validity of the method and show significant improvements over the state-of-the-art in velocity estimation and accumulation of radar points.
29	Proceedings of the IEEE ICCV, 2021, pp.10996-11005	EventHPE: Event-Based 3D Human Pose and Shape Estimation	Shihao Zou, Chuan Guo, Xinxin Zuo, Sen Wang, Pengyu Wang, Xiaoqin Hu, Shoushun Chen, Minglun Gong, Li Cheng	University of Alberta	相关(CNN)	Event camera is an emerging imaging sensor for capturing dynamics of moving objects as events, which motivates our work in estimating 3D human pose and shape from the event signals. Events, on the other hand, have their unique challenges: rather than capturing static body postures, the event signals are best at capturing local motions. This leads us to propose a two-stage deep learning approach, called EventHPE. The first-stage, FlowNet, is trained by unsupervised learning to infer optical flow from events. Both events and optical flow are closely related to human body dynamics, which are fed as input to the ShapeNet in the second stage, to estimate 3D human shapes. To mitigate the discrepancy between image-based flow (optical flow) and shape-based flow (vertices movement of human body shape), a novel flow coherence loss is introduced by exploiting the fact that both flows are originated from the identical human motion. An in-house event-based 3D human dataset is curated that comes with 3D pose and shape annotations, which is by far the largest one to our knowledge. Empirical evaluations on DHP19 dataset and our in-house dataset demonstrate the effectiveness of our approach.
30	Proceedings of the IEEE ICCV, 2021, pp.16228-16237	CTRL-C: Camera Calibration TRansformer With Line-Classification	Jinwoo Lee, Hyunsung Go, Hyunjoon Lee, Sunghyun Cho, Minhyuk Sung, Junho Kim	Kakao Brain	相关(CNN)	Single image camera calibration is the task of estimating the camera parameters from a single input image, such as the vanishing points, focal length, and horizon line. In this work, we propose Camera calibration TRansformer with Line-Classification (CTRL-C), an end-to-end neural network-based approach to single image camera calibration, which directly estimates the camera parameters from an image and a set of line segments. Our network adopts the transformer architecture to capture the global structure of an image with multi-modal inputs in an end-to-end manner. We also propose an auxiliary task of line classification to train the network to extract the global geometric information from lines effectively. Our experiments demonstrate that CTRL-C outperforms the previous state-of-the-art methods on the Google Street View and SUN360 benchmark datasets. Code is available at https://github.com/jwlee-vcl/CTRL-C.
31	Proceedings of the IEEE ICCV, 2021, pp.2135-2145	Dual Transfer Learning for Event-Based End-Task Prediction via Pluggable Event to Image Translation	Lin Wang, Yujeong Chae, Kuk-Jin Yoon	Visual Intelligence Lab., KAIST, Korea	相关(DNN)	Event cameras are novel sensors that perceive the per-pixel intensity changes and output asynchronous event streams with high dynamic range and less motion blur. It has been shown that events alone can be used for end-task learning, e.g., semantic segmentation, based on encoder-decoder-like networks. However, as events are sparse and mostly reflect edge information, it is difficult to recover original details merely relying on the decoder. Moreover, most methods resort to the pixel-wise loss alone for supervision, which might be insufficient to fully exploit the visual details from sparse events, thus leading to less optimal performance. In this paper, we propose a simple yet flexible two-stream framework named Dual Transfer Learning (DTL) to effectively enhance the performance on the end-tasks without adding extra inference cost. The proposed approach consists of three parts: event to end-task learning (EEL) branch, event to image translation (EIT) branch, and transfer learning (TL) module that simultaneously explores the feature-level affinity information and pixel-level knowledge from the EIT branch to improve the EEL branch. This simple yet novel method leads to strong representation learning from events and is evidenced by the significant performance boost on the end-tasks such as semantic segmentation and depth estimation.
32	Proceedings of the IEEE ICCV, 2021, pp.14528-14538	EgoRenderer: Rendering Human Avatars From Egocentric Camera Images	Tao Hu, Kripasindhu Sarkar, Lingjie Liu, Matthias Zwicker, Christian Theobalt	Department of Computer Science, University of Maryland, College Park	不相关	We present EgoRenderer, a system for rendering full-body neural avatars of a person captured by a wearable, egocentric fisheye camera that is mounted on a cap or a VR headset. Our system renders photorealistic novel views of the actor and her motion from arbitrary virtual camera locations. Rendering full-body avatars from such egocentric images come with unique challenges due to the top-down view and large distortions. We tackle these challenges by decomposing the rendering process into several steps, including texture synthesis, pose construction, and neural image translation. For texture synthesis, we propose Ego-DPNet, a neural network that infers dense correspondences between the input fisheye images and an underlying parametric body model, and to extract textures from egocentric inputs. In addition, to encode dynamic appearances, our approach also learns an implicit texture stack that captures detailed appearance variation across poses and viewpoints. For correct pose generation, we first estimate body pose from the egocentric view using a parametric model. We then synthesize an external free-viewpoint pose image by projecting the parametric model to the user-specified target viewpoint. We next combine the target pose image and the textures into a combined feature image, which is transformed into the output color image using a neural image translation network. Experimental evaluations show that EgoRenderer is capable of generating realistic free-viewpoint avatars of a person wearing an egocentric camera. Comparisons to several baselines

	33	Proceedings of the IEEE ICCV, 2021, pp.10221-10230	Temporal-Wise Attention Spiking Neural Networks for Event Streams Classification	Man Yao, Huanhuan Gao, Guangshe Zhao, Dingheng Wang, Yihan Lin, Zhaoxu Yang, Guoqi Li	Xi'an Jiaotong University	相关(SNN)	How to effectively and efficiently deal with spatio-temporal event streams, where the events are generally sparse and non-uniform and have the us temporal resolution, is of great value and has various real-life applications. Spiking neural network (SNN), as one of the brain-inspired event-triggered computing models, has the potential to extract effective spatio-temporal features from the event streams. However, when aggregating individual events into frames with a new higher temporal resolution, existing SNN models do not attach importance to that the serial frames have different signal-to-noise ratios since event streams are sparse and non-uniform. This situation interferes with the performance of existing SNNs. In this work, we propose a temporal-wise attention SNN (TA-SNN) model to learn frame-based representation for processing event streams. Concretely, we extend the attention concept to temporal-wise input to judge the significance of frames for the final decision at the training stage, and discard the irrelevant frames at the inference stage. We demonstrate that TA-SNN models improve the accuracy of event streams classification tasks. We also study the impact of multiple-scale temporal resolutions for frame-based representation. Our approach is tested on three different classification tasks: gesture recognition, image classification, and spoken digit recognition. We report the state-of-the-art results on these tasks, and get the essential improvement of accuracy (almost 19%) for gesture recognition with only 60 ms.