# Autonomous Vehicle Vision 2021: ICCV Workshop Summary

Rui Fan[1], Nemanja Djuric[2], Fisher Yu[3], Rowan McAllister[4], Ioannis Pitas[5]
[1]Tongji University, P. R. China [2]Aurora Innovation, USA [3]ETH Zürich, Switzerland
[4]Toyota Research Institute, USA [5]Aristotle University of Thessaloniki, Greece
rui.fan@ieee.org, nemanja@temple.edu, i@yf.io,
rowan.mcallister@tri.global, pitas@csd.auth.gr

## Abstract

*This paper summarizes the 2nd Autonomous Vehicle Vision (AVVision) workshop (avvision.xyz/iccv21), organized virtually in conjunction with ICCV 2021. The organizers invited seven experts from both industry and academia to deliver keynote talks, discussing the state-of-the-art and challenges in the field of autonomous driving. A total of 27 papers were accepted for publication in the ICCV 2021 proceedings (IEEE Xplore and CVF open access), resulting in an acceptance rate of* 50.9%. *In addition to serving as a workshop summary and a brief overview of the existing challenges, this paper also presents how these challenges were addressed by the authors through their proposed solutions.*

## 1. Introduction

Due to the recent boom in artificial intelligence technologies, there are growing expectations that fully autonomous driving may become a reality in the near future. This is expected to bring fundamental changes to our society, as fully autonomous vehicles offer great potential to improve efficiency on roads, reduce traffic accidents, increase productivity, and minimize our environmental impact. Ensuring safety of autonomous vehicles requires a multi-disciplinary approach across all the levels of functional hierarchy, from hardware fault tolerance to modern machine learning, over cooperation with humans driving conventional vehicles and system validation for operations in highly unstructured environments, to appropriate regulatory approaches.

As a key component of autonomous driving, autonomous vehicle vision systems are typically developed based on cutting-edge computer vision, machine learning, image and signal processing, and advanced sensing technologies. With recent advances in deep learning, autonomous vehicle vision systems have achieved very compelling results. However, there still exist many challenges. For instance, the perception modules cannot perform well in poor weather and/or illumination conditions, or in complex urban environments. Developing robust and all-weather perception algorithms is a popular research area that requires more attention. In addition, most perception methods are generally computationally-intensive and cannot run in real-time on embedded and resource-limited hardware. Therefore, fully exploiting the parallel-computing architecture, such as embedded GPUs, for real-time perception, prediction, and planning is a hot topic that is researched in the autonomous driving field. Furthermore, existing supervised learning approaches have achieved strong results, but their performance is fully dependent on the quality and amount of labeled training data. Labeling such data is a time-consuming and labor-intensive process. Unsupervised and self-supervised learning approaches and domain adaptation techniques are, therefore, becoming increasingly crucial for real-world autonomous driving applications.

## 2. AVVision Workshop

Autonomous Vehicle Vision (AVVision) (avvision.xyz) serves as a premier platform and foundation for the technology of tomorrow. In January 2021, AVVision organized the 1st Workshop at WACV 2021 (avvision.xyz/wacv21), where the organizers invited four keynote speakers to deliver talks, had eight peer-reviewed paper presentations, and hosted three competitions. The first AVVision workshop attracted wide attention from the autonomous driving community. Following the success of the 1st AVVision workshop, we aim to host the AVVision workshop annually to share knowledge about cutting-edge technologies, discuss the key and difficult problems, and propose effective solutions to existing challenges.

This paper summarizes the 2nd AVVision workshop organized in conjunction with ICCV 2021 (avvision.xyz/iccv21), to be held virtually on October 17, 2021. The workshop aims to bring together industry professionals and academics to brainstorm and exchange ideas on the

*All authors contributed equally to this work.

advancement of autonomous vehicle vision applications to help cope with the aforementioned challenges. The rest of this paper summarizes the 2nd AVVision workshop and introduces the accepted papers.

## 3. Keynote Talks

We invited seven speakers to deliver keynote talks, including Cordelia Schmid from INRIA, Raquel Urtasun from the University of Toronto, Andreas Geiger from the University of Tübingen, Fisher Yu from ETH Zürich, Laura Leal-Taixé from the Technical University of Munich, Matthew Johnson-Roberson from the University of Michigan, and Carl Wellington from Aurora Innovation. Their keynote talks will be made available at the AVVision YouTube channel (youtube.com/playlist?list=PL51IczHEeqwSdk2U7Bx9ev-bmkdkzjOtf).

## 4. Contributed Papers

We received 53 regular paper submissions, which discussed existing challenges and the state-of-the-art (SoTA) in autonomous driving. Each submission was reviewed by at least three PC members and/or chairs. 27 papers were ultimately accepted for publication at the AVVision workshop, resulting in the acceptance rate of 50.9%. The organizers and PC members nominated one submission to receive the best paper award, sponsored by the Nvidia Corporation. In the remainder of this section, we briefly introduce each accepted paper. The full papers are publicly available at the CVF open access website (openaccess.thecvf.com).

**SA-Det3D: Self-Attention Based Context-Aware 3D Object Detection [1]** In this paper, the authors proposed two variants of self-attention for contextual modeling in 3D object detection by augmenting convolutional features with self-attention features. They first incorporated the pairwise self-attention mechanism into the current state-of-the-art BEV-, voxel-, and point-based detectors. On the KITTI validation set, they showed consistent improvement over strong baseline models of up to 1.5 3D AP while simultaneously reducing their parameter footprint and computational cost by 15-80% and 30-50%, respectively. They also proposed a self-attention variant that samples a subset of the most representative features by learning deformations over randomly sampled locations. This allows them to scale explicit global contextual modeling to larger point clouds and leads to more discriminative and informative feature descriptors. Their method can be flexibly applied to most SoTA detectors with increased accuracy and parameter and compute efficiency. They demonstrated that their proposed method improves 3D object detection performance on KITTI, nuScenes, and Waymo Open data sets.

**Visual Reasoning Using Graph Convolutional Networks for Predicting Pedestrian Crossing Intention [2]** In this paper, the authors employed rich visual features in graph convolutional auto-encoders to encode the relationship between the pedestrian and its surrounding objects to reason their crossing intention. They also incorporated pedestrian bounding boxes and human pose estimation in the prediction module to further improve prediction results. It was demonstrated that their proposed model outperforms the SoTA methods by a wide margin, particularly for cases where the pedestrian has no crossing intention.

**Cross-modal Matching CNN for Autonomous Driving Sensor Data Monitoring [3]** In this paper, the authors presented a cross-modal convolutional neural network (CNN) for autonomous driving sensor data monitoring functions, such as fault detection and online data quality assessment. Assuming the overlapping view of different sensors should be consistent under normal circumstances, the authors detected anomalies such as mis-synchronization through matching camera images and LIDAR point clouds. A masked pixel-wise metric learning loss was proposed to improve exploration of local structures and build an alignment-sensitive pixel embedding. In their experiments with a selected KITTI data set and specially tailored fault data generation methods, their approach showed promising success for sensor fault detection and point cloud quality assessment results.

**A Computer Vision-Based Attention Generator Using Reinforcement Learning [4]** This paper presented an end-to-end computer vision-based reinforcement learning (RL) technique that intelligently selects a priority region of an image to place greater attention in order to achieve better perception performance. This method was evaluated on the Berkeley Deep Drive (BDD) dataset. Results demonstrated that, compared to a baseline method, a substantial improvement in perception performance can be attained at a minimal cost in terms of time and processing.

**Semantics-Aware Multi-Modal Domain Translation: From LiDAR Point Clouds to Panoramic Color Images [5]** In this work, the authors presented a simple yet effective framework to address the domain translation problem between different sensor modalities with unique data formats. By relying only on the semantics of the scene, their modular generative framework can, for the first time, synthesize a panoramic color image from a given full 3D LiDAR point cloud. The framework starts with semantic segmentation of the point cloud, which was initially projected onto a spherical surface. The same semantic segmentation was applied to the corresponding camera image. Next, their new conditional generative model adversarially learned to

translate the predicted LiDAR segment maps to the camera image counterparts. Finally, generated image segments were processed to render the panoramic scene images. The authors provided a thorough quantitative evaluation on the SemanticKitti data set and show that our proposed framework outperforms other strong baseline models.

**SCARF: A Semantic Constrained Attention Refinement Network for Semantic Segmentation [6]** In this paper, the authors proposed an end-to-end semantic constrained attention refinement (SCARF) network to fully utilize the semantic information across different layers based on semantic constrained contextual dependencies. Their novelties lie in the following aspects: 1) they presented a general framework for capturing the non-local contextual dependencies; 2) within the framework, they introduced an efficient category attention (CA) block to capture semantic-related context by using the category constraint from coarse segmentation, which reduces the computational complexity from $O(n^2)$ to $O(n)$ for image with $n$ pixels; 3) they overcame the contextual information confusion problem by adapting a category-wise learning weight to balance the non-local contextual dependencies and the local consistency problem; 4) they fully utilized the multi-scale semantic-related contextual information by refining the segmentation iteratively across layers with semantic constraints. Extensive evaluations demonstrated that their proposed SCARF network significantly improves the segmentation results and achieves superior performance on PASCAL VOC 2012, PASCAL Context, and Cityscapes data sets.

**SDVTracker: Real-Time Multi-Sensor Association and Tracking for Self-Driving Vehicles [7]** This paper presented a practical and lightweight tracking system, SDVTracker, that uses a deep model for association and state estimation in conjunction with an interacting multiple model (IMM) filter. The proposed tracking method is fast, robust, and generalizes across multiple sensor modalities and different Vulnerable Road User (VRU) classes. The authors detailed a model that jointly optimizes both association and state estimation with a novel loss, an algorithm for determining ground-truth supervision, and a training procedure. They showed that the proposed system significantly outperforms hand-engineered methods on a real-world urban driving data while running in less than 2.5 ms on CPU for a scene with 100 actors, making it suitable for self-driving vehicles where low latency and high accuracy are critical.

**Speak2Label: Using Domain Knowledge for Creating a Large Scale Driver Gaze Zone Estimation Dataset [8]** In this paper, a fully automated technique for labeling an image-based gaze behavior data set for driver gaze zone estimation was proposed. Domain knowledge was added to the data recording paradigm, and later labels were generated in an automated manner using Speech-To-Text conversion (STT). To remove the noise in the STT process due to different illumination and ethnicity of subjects in the data, the speech frequency and energy were analyzed. The resultant Driver Gaze in the Wild (DGW) data set contains 586 recordings captured during different times of the day, including evenings. The large-scale data set contains 338 subjects with an age range of 18-63 years. As the data were recorded in different lighting conditions, an illumination-robust layer was proposed in the CNN. The extensive experiments show the variance in the data set resembling real-world conditions and the effectiveness of the proposed CNN pipeline. The proposed network was also fine-tuned for the eye gaze prediction task, which shows the discriminativeness of the representation learned by their network on the proposed DGW data set.

**DriPE: A Dataset for Human Pose Estimation in Real-World Driving Settings [9]** This paper first introduced Driver Pose Estimation (DriPE), a new data set to foster the development and evaluation of methods for human pose estimation of drivers in consumer vehicles. This is claimed to be the first publicly available data set depicting drivers in real scenes. It contains 10k images of 19 different driver subjects, manually annotated with human body keypoints and an object bounding box. Furthermore, this paper proposed a new keypoint-based metric for human pose estimation. This metric highlights the limitations of current metrics for HPE evaluation and the limitations of existing deep neural networks on pose estimation, both on general and driving-related data sets.

**Efficient Uncertainty Estimation in Semantic Segmentation via Distillation [10]** Inspired by the concept of knowledge distillation, whereby the performance of a compact model is improved by training it to mimic the outputs of a larger model, the authors trained a compact model to mimic the output distribution of a large ensemble of models, such that for each output there is a prediction and a predicted level of uncertainty for that prediction. They applied uncertainty distillation in the context of a semantic segmentation task for autonomous vehicle scene understanding and demonstrated a capability to predict pixel-wise uncertainty over the resultant class probability map reliably. They also showed that the aggregate pixel uncertainty across an image can be used as a metric for reliable detection of out-of-distribution data.

**RaidaR: A Rich Annotated Image Dataset of Rainy Street Scenes [11]** This paper introduced RaidaR, a rich annotated image data set of rainy street scenes, to support autonomous driving research. The new data set contains

the largest number of rainy images (58,542) to date, 5,000 of which provide semantic segmentation annotations and 3,658 provide object instance segmentation annotations. The RaidaR images cover a wide range of realistic rain-induced artifacts, including fog, droplets, and road reflections, which can effectively augment existing street scene data sets to improve data-driven machine perception during rainy weather. To facilitate efficient annotation of a large volume of images, the authors developed a semi-automatic scheme combining manual segmentation and an automated processing akin to cross validation, resulting in a 10-20 fold reduction in annotation time. They demonstrated the utility of their new data set by showing how data augmentation with RaidaR can elevate the accuracy of existing segmentation algorithms. They also presented a novel unpaired image-to-image translation algorithm for adding/removing rain artifacts, which directly benefits from RaidaR.

**Synthetic Data Generation using Imitation Training [12]** The authors proposed a strategic approach to generate synthetic data to improve machine learning algorithms. The utilization of synthetic data has shown promising results yet there are no specific rules or recipes on how to generate and cook synthetic data. They proposed imitation training as a guideline of synthetic data generation to add more underrepresented entities and balance the data distribution for networks to handle corner cases and resolve long-tail problems. The proposed imitation training has a circular process with three main steps: 1) the existing system was evaluated, and failure cases such as false positive and false negative detections were sorted out; 2) synthetic data imitating such failure cases were created with domain randomization; 3) they trained a network with the existing data and the newly added synthetic data; They repeated these three steps until the evaluation metric converges. They validated the approach by experimenting with respect to object detection in autonomous driving.

**SS-SFDA : Self-Supervised Source-Free Domain Adaptation for Road Segmentation in Hazardous Environments [13]** This paper presented a novel approach for unsupervised road segmentation in adverse weather conditions like rain or fog. This includes a new algorithm for source-free domain adaptation (SFDA) using self-supervised learning. Moreover, their approach used several techniques to address various challenges in SFDA and improve performance, including online generation of pseudo-labels and self-attention and use of curriculum learning, entropy minimization, and model distillation. The authors evaluated the performance on six data sets corresponding to real and synthetic adverse weather conditions. Their proposed method outperforms all prior works on unsupervised road segmentation and SFDA by at least 10.26%. Moreover, their self-supervised algorithm exhibits similar accuracy performance in terms of mIOU score compared to previous supervised methods.

**YOLinO: Generic Single Shot Polyline Detection in Real Time [14]** The authors proposed an approach that builds upon the idea of single-shot object detection. Reformulating the problem of polyline detection as a bottom-up composition of small line segments allows detecting bounded, dashed, and continuous polylines with a single head. This has several major advantages over previous methods. Not only is the method at 187 fps more than suited for real-time applications with virtually any restriction on the shapes of the detected polylines. By predicting multiple line segments for each cell, even branching or crossing polylines can be detected. They evaluated their approach on three different applications for road marking, lane border, and centerline detection. Hereby, they demonstrated the ability to generalize to different domains and both implicit and explicit polyline detection tasks.

**Graph Convolutional Networks for 3D Object Detection on Radar Data [15]** In this work, the authors focused on fully leveraging raw radar tensor data instead of building up on human-biased point clouds, which are typical of traditional radar signal processing techniques. Utilizing a graph neural network on the raw radar tensor, they significantly improved +10% in average precision over a grid-based convolutional baseline network. The performance of both networks was evaluated on a real-world data set with dense city traffic scenarios, diverse object orientations and distances, and occlusions up to visually fully occluded objects. Their proposed network increases the maximum range for SoTA full-3D object detection on radar data from 20m to 100m.

**Multi-Weather City: Adverse Weather Stacking for Autonomous Driving [16]** Based on GAN and CycleGAN architectures, the authors proposed an overall (modular) architecture for constructing data sets, allowing one to add, swap out, and combine components to generate images with diverse weather conditions. Starting from a single (overcast) data set with ground-truth, they generate seven versions of the same data in diverse weather and proposed an extension to augment the generated conditions, thus resulting in a total of 14 adverse weather conditions, requiring a single ground truth. They tested the generated conditions' quality in terms of perceptual quality and suitability for training downstream tasks, using real-world, out-of-distribution adverse weather extracted from various data sets. They showed improvements in object detection and instance segmentation across all conditions, in many cases exceeding 10% increase in AP, and provided the materials

and instructions needed to re-construct the multi-weather data set, based upon the original Cityscapes data set.

**Frustum-PointPillars: A Multi-Stage Approach for 3D Object Detection using RGB Camera and LiDAR [17]** The authors proposed a novel method called Frustum-PointPillars for 3D object detection using LiDAR data. Instead of solely relying on point cloud features, they leveraged the mature field of 2D object detection to reduce the search space in the 3D space. Then, they used the Pillar Feature Encoding network for object localization in the reduced point cloud. They also proposed a novel approach for masking point clouds to improve the localization of objects further. They train their network on the KITTI data set and perform experiments to show the effectiveness of our network. On the KITTI test set, their method outperforms other SoTA multi-sensor approaches for localization of Pedestrians in 3D (BEV detection). Their method achieves a runtime of 14 Hz, which is significantly faster than other multi-sensor SoTA approaches.

**CenterPoly: Real-Time Instance Segmentation Using Bounding Polygons [18]** The authors presented a novel method, called CenterPoly, for real-time instance segmentation using bounding polygons. They applied it to detect road users in dense urban environments, making it suitable for applications in intelligent transportation systems like automated vehicles. CenterPoly detects objects by their center keypoint while predicting a fixed number of polygon vertices for each object, thus performing detection and segmentation in parallel. The network heads share most of the network parameters, making it fast and lightweight enough to run at in real time. To properly convert mask ground-truth to polygon ground-truth, the authors designed a vertex selection strategy to facilitate the learning of the polygons. Additionally, to better segment overlapping objects in dense urban scenes, the authors also trained a relative depth branch to determine which instances are closer and which are further, using available weak annotations. They proposed several models with different backbones to show the possible speed/accuracy trade-offs. The models were trained and evaluated on Cityscapes, KITTI, and IDD, and the achieved SoTA results were published on these public benchmarks.

**Occupancy Grid Mapping with Cognitive Plausibility for Autonomous Driving Applications [19]** This work investigated the validity of an occupancy grid mapping inspired by human cognition and how humans visually perceive the environment. This query is motivated by the fact that, to date, no autonomous driving system reaches the performance of an ordinary human driver. The mechanisms behind human perception could provide cues on how to improve common techniques employed in autonomous navigation, specifically occupancy grids to represent the environment. The authors experimented with a neural network that maps an image of the scene onto an occupancy grid representation. They showed how the model benefits from two key (and yet simple) changes: 1) a different format of occupancy grid that resembles the way the brain projects the environment into a warped representation in the cortical visual area; 2) a mechanism similar to human visual attention that filters out non-relevant information from the scene. These effective expedients can potentially be applied to any autonomous driving task requiring a scenario abstract representation like the occupancy grids.

**It's All Around You: Range-Guided Cylindrical Network for 3D Object Detection [20]** This work presented a novel approach for analyzing 3D data produced by 360-degree depth scanners, utilizing a more suitable coordinate system aligned with the scanning pattern. Furthermore, the authors introduced a novel notion of range-guided convolutions, adapting the receptive field by distance from the ego vehicle and the object's scale. Their developed network demonstrated impressive results on the competitive nuScenes 3D object detection challenge, comparable to current SoTA architectures.

**Causal BERT: Improving Object Detection by Searching for Challenging Groups [21]** The first contribution of this paper is a method to find such groups in foresight, leveraging advances in simulation and masked language modeling to perform causal interventions on simulated driving scenes. The authors then used the found groups to improve detection, exemplified by Diamondback bikes, whose performance we improve by 30 AP points. Such a solution is of high priority because it would significantly enhance the robustness and safety of AV systems. The second contribution of this paper is the tooling to run interventions, which benefits the causal community tremendously.

**On the Road to Large-Scale 3D Monocular Scene Reconstruction using Deep Implicit Functions [22]** In this work, the authors mainly tested the limits of deep implicit functions by applying them to the task of reconstructing large-scale, real-world traffic scenes from a single monocular image. Since watertight meshes are not in general available for real-world scenes, they proposed an alternative training scheme using LiDAR to provide approximate ground truth occupancy supervision. They also demonstrated that incorporating priors such as pre-detected object bounding boxes can vastly improve reconstruction quality. They demonstrated their method on a large-scale, real-world autonomous driving data set.

**Weakly Supervised Approach for Joint Object and Lane Marking Detection [23]** This paper proposed a weakly supervised approach for joint object and lane marking detection. Given an image from the lane marking detection data set, they used a pre-trained network to label different objects within a scene, generating pseudo bounding boxes used to train a network that jointly detects objects and lane markings. With an emphasis on inference speed and performance, they proposed two architectures based on CNNs and Transformers. The CNN-based approach uses row-based pixel classification to detect and cluster lane markings alongside a single-stage anchor free object detector while sharing the same encoder backbone. Alternatively, using dual decoders, the transformer-based approach directly estimates bounding boxes and polynomial coefficients of lane markings. Through extensive qualitative and quantities experiments, they demonstrated the efficacy of the proposed architectures on leading data sets for object and lane marking detection and report SoTA performance per GFLOPs.

**Few-Shot Batch Incremental Road Object Detection via Detector Fusion [24]** The authors tackled batch incremental few-shot road object detection using data from the India Driving Dataset (IDD). Their approach, DualFusion, combined object detectors to learn to detect rare objects with very limited data, all without severely degrading the detector's performance on the abundant classes. In the IDD OpenSet incremental few-shot detection task, their approach achieved an mAP50 score of 40.0 on the base classes and an overall mAP50 score of 38.8, both of which are the highest to date. In the COCO batch incremental few-shot detection task, they achieved a novel AP score of 9.9, surpassing the SoTA novel class performance by over $6.6\times$.

**Multistage Fusion for Multi-Class 3D Lidar Detection [25]** This paper proposed a LiDAR-camera fusion method for multi-class 3D object detection. It makes the utmost use of data from the two sensors by multiple fusion stages and can be learned in an end-to-end manner. First, the authors applied a multi-level gated adaptive fusion mechanism with the feature extraction backbone. This point-wise fusion stage assiduously exploits the image and point cloud inputs and obtains joint semantic representations of the scene. Next, given the regions of interest proposed based on the LiDAR features, the corresponding Camera features are selected by RoI-based feature pooling. These features were used to enrich the LiDAR features in local regions and enhance the proposal refinement. Moreover, they introduced a multi-label classification task as an additional regularization to the object detection network. Without relying on extra labels, it helps the model better mine the extracted features and discover hard object instances. The experiments conducted on the KITTI data set demonstrated that all their proposed fusion strategies are effective.

**CDAda: A Curriculum Domain Adaptation for Nighttime Semantic Segmentation [26]** In this paper, the authors proposed a curriculum domain adaptation method (CDAda) to realize smooth semantic knowledge transfer from daytime to nighttime. CDAda consists of two steps: 1) inter-domain style adaptation: fine-tune the daytime-trained model on the labeled synthetic nighttime images through the proposed frequency-based style transformation method (replace the low-frequency components of daytime images with those of nighttime images); 2) intra-domain gradual self-training: separate the nighttime domain into the easy split nighttime domain and hard split nighttime domain based on the "entropy + illumination" ranking principle, then gradually adapt the model to the two sub-domains through pseudo supervision on easy split data and entropy minimization on hard split data. Extensive experiments on the Nighttime Driving, Dark Zurich, and BDD100K-night data sets demonstrated the effectiveness of their approach against existing SoTA approaches.

**Monocular 3D Localization of Vehicles in Road Scenes [27]** This paper proposed a monocular vision-based framework for 3D-based detection, tracking, and localization by effectively integrating all three tasks in a complementary manner. Their system contains an RCNN-based Localization Network (LOCNet), which works with fitness evaluation score (FES)-based single-frame optimization to achieve more accurate and refined 3D vehicle localization. To better utilize the temporal information, they further used a multi-frame optimization technique, taking advantage of camera ego-motion and a 3D TrackletNet Tracker (3D TNT), to improve accuracy and consistency in the 3D localization results. Their system outperforms SoTA image-based solutions in diverse scenarios and is even comparable with LiDAR-based methods.

## 5. Organizers

**Rui (Ranger) Fan, Tongji University** Rui Fan received the B.Eng. degree in automation (control science & engineering) from the Harbin Institute of Technology, China, in 2015 and the Ph.D. degree in electrical and electronic engineering from the University of Bristol, U.K., in 2018. From 2018 to 2020, he was a Research Associate with the Robotics Institute and the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong. Between 2020 and 2021, he was a Postdoctoral Fellow with the Department of Ophthalmology and the Department of Computer Science and Engineering, University of California San Diego, USA. He is currently a (full) Research Professor with the Department

of Control Science and Engineering, College of Electronics and Information Engineering, and Shanghai Research Institute for Intelligent Autonomous Systems, Tongji University, Shanghai, China.

**Nemanja Djuric, Aurora Technologies** Nemanja is a Principal Software Engineer and Technical Lead Manager at Aurora Technologies, while prior to his current position he worked in a similar role at Uber ATG. Previously he was a Research Scientist at Yahoo Labs working on computational advertising. Dr. Djuric published more than 50 peer-reviewed publications at leading Machine Learning, Data Mining, Computer Vision, and Web Science conferences and journals, in addition to 4 granted patents and 10+ pending patent applications. His work was featured in Market Watch, VentureBeat, IEEE Innovation at Work, and other news outlets across the world. He has extensive experience in organizing workshops at top-tier venues on a variety of topics, including AdKDD and TargetAd workshops from 2015 until today (organized in conjunction with the KDD, WWW, and WSDM conferences), as well as "Precognition: Seeing through the Future" workshop organized in conjunction with CVPR in 2019, 2020, and 2021.

**Fisher Yu, ETH Zürich** Fisher Yu is an Assistant Professor at ETH Zürich in Switzerland. He obtained the Ph.D. degree from Princeton University and became a postdoctoral researcher at UC Berkeley. He directs the Visual Intelligence and Systems (VIS) Group in the Computer Vision Lab. His goal is to build perceptual systems capable of performing complex tasks in complex environments. His research is at the junction of machine learning, computer vision, and robotics. He currently works on closing the loop between vision and action. He has rich experience in organizing autonomous driving workshops at top-tier machine learning and CV conferences. He is the co-organizer of Workshops on Autonomous Driving at CVPR in 2017, 2018, 2019, 2020, 2021, and Workshop on Machine Learning for Autonomous Driving at NeurIPS in 2019, 2020.

**Rowan McAllister, Toyota Research Institute** Rowan is a Research Scientist at Toyota Research Institute in California. He obtained his PhD degree in machine learning from the University of Cambridge, and was a postdoctoral researcher at UC Berkeley working on autonomous vehicles. His expertise is in using machine learning for autonomous vehicle planning. His prior organizational experience includes the NeurIPS Workshop on Machine Learning for Autonomous Driving (2019, 2020), ECCV Workshop on Perception for Autonomous Driving (2020), ICML Workshop on AI for Autonomous Driving (2020), and the RSS Workshop on Interaction and Decision-Making in Autonomous Driving (2020).

**Ioannis Pitas, Aristotle University of Thessaloniki** Ioannis Pitas (IEEE fellow) received his Diploma and Ph.D. degree in Electrical Engineering, both from the Aristotle University of Thessaloniki (AUTH), Greece. Since 1994, he has been a Professor at the Department of Informatics of AUTH and Director of the Artificial Intelligence and Information Analysis (AIIA) lab. He served as a Visiting Professor at several Universities. Prof. Pitas has published over 1000 papers, contributed to 47 books in his areas of interest and edited or (co-)authored another 11 books. He has also been a member of the program committee of many scientific conferences and workshops. In the past he served as Associate Editor or co-Editor of 9 international journals and General or Technical Chair of 4 international conferences. He participated in 71 R&D projects, primarily funded by the European Union and is/was principal investigator in 42 such projects. Prof. Pitas leads the big European H2020 R&D project MULTIDRONE: multidrone.eu. He is AUTH principal investigator in H2020 R&D projects Aerial Core and AI4Media. He is chair of the Autonomous Systems Initiative ieeeasi.signalprocessingsociety.org. He is head of the EC funded AI doctoral school of Horizon2020 EU funded R&D project AI4Media. Prof. Pitas' current interests are in the areas of computer vision, machine learning, autonomous systems, and image/video processing.

## References


[1] Prarthana Bhattacharyya et al. Sa-det3d: Self-attention based context-aware 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 2

[2] Tina Chen et al. Visual reasoning using graph convolutional networks for predicting pedestrian crossing intention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 2

[3] Yiqiang Chen et al. Cross-modal matching cnn for autonomous driving sensor data monitoring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 2

[4] Jordan Chipka et al. A computer vision-based attention generator using reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 2

[5] Tiago Cortinhal et al. Semantics-aware multi-modal domain translation: From lidar point clouds to panoramic color images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 2

[6] Xiaofeng Ding et al. Scarf: A semantic constrained attention refinement network for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 3

[7] Shivam Gautam et al. Sdvtracker: Real-time multi-sensor association and tracking for self-driving vehicles. In *Proceed-*

*ings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 3

[8] Shreya Ghosh et al. Speak2label: Using domain knowledge for creating a large scale driver gaze zone estimation dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 3

[9] Romain Guesdon et al. Dripe: A dataset for human pose estimation in real-world driving settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 3

[10] Christopher Holder and Muhammad Shafique. Efficient uncertainty estimation in semantic segmentation via distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 3

[11] Jiongchao Jin et al. Raidar: A rich annotated image dataset of rainy street scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 3

[12] Aman Kishore et al. Synthetic data generation using imitation training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 4

[13] Divya Kothandaraman et al. Ss-sfda : Self-supervised source-free domain adaptation for road segmentation in hazardous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 4

[14] Annika Meyer et al. Yolino: Generic single shot polyline detection in real time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 4

[15] Michael Meyer et al. Graph convolutional networks for 3d object detection on radar data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 4

[16] Valentina Musat et al. Multi-weather city: Adverse weather stacking for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 4

[17] Anshul Paigwar et al. Frustum-pointpillars: A multi-stage approach for 3d object detection using rgb camera and lidar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 5

[18] Hughes Perreault et al. Centerpoly: Real-time instance segmentation using bounding polygons. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 5

[19] Alice Plebe et al. Occupancy grid mapping with cognitive plausibility for autonomous driving applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 5

[20] Meytal Rapoport-Lavie and Dan Raviv. It's all around you: Range-guided cylindrical network for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 5

[21] Cinjon Resnick et al. Causal bert: Improving object detection by searching for challenging groups. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 5

[22] Thomas Roddick et al. On the road to large-scale 3d monocular scene reconstruction using deep implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 5

[23] Pranjay Shyam et al. Weakly supervised approach for joint object and lane marking detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 6

[24] Anuj Tambwekar et al. Few-shot batch incremental road object detection via detector fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 6

[25] Zejie Wang et al. Multistage fusion for multi-class 3d lidar detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 6

[26] Qi Xu et al. Cdada: A curriculum domain adaptation for nighttime semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 6

[27] Haotian Zhang et al. Monocular 3d localization of vehicles in road scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, October 2021. 6