Scientific report

on project implementation status between January-December 2020

Proiect PN-III-P4-ID-PCCF-2016-0180

Financing: UEFISCDI, Nr. 9/2018

Project homepage: http://www.imar.ro/clvp/projects.php?ID_p=15

Integrated Semantic Visual Perception and Control for Autonomous Systems

"Simion Stoilow" Institute of Mathematics of the Romanian Academy (IMAR)

Principal Investigator (PI): Prof-univ. dr. Cristian Sminchisescu

Co-PI: Prof-univ. Dr. Sergiu Nedevschi

The third year of the project continued to follow the plan of action, in partnership with the research group of Prof. Sergiu Nedevschi from the Technical University of Cluj-Napoca. This report convers the achievements of this period and are highlighted below.

A1. Deep 3D Reconstruction

Task 1.2: Deep Structured Geometric Models with Semantics

Generating good quality and geometrically plausible synthetic images of humans with the ability to control appearance, pose and shape parameters, has become increasingly important for a variety of tasks ranging from photo editing, fashion virtual try-on, to special effects and image compression. In this project, we propose a HUSC (HUman Synthesis and Scene Compositing) framework for the realistic synthesis of humans with different appearance, in novel poses and scenes. Central to our formulation is 3d reasoning for both people and scenes, in order to produce realistic collages, by correctly modeling perspective effects and occlusion, by taking into account scene semantics and by adequately handling relative scales. Conceptually our framework consists of three components: (1) a human image synthesis model with controllable pose and appearance, based on a parametric representation, (2) a person insertion procedure that leverages the geometry and semantics of the 3d scene, and (3) an appearance compositing process to create a seamless blending between the colors of the scene and the generated human image, and avoid visual artifacts. The performance of our framework is supported by both qualitative and quantitative results, in particular state-of-the art synthesis scores for the DeepFashion dataset. Qualitative results of our method can be seen below, in fig. T1.2.2.

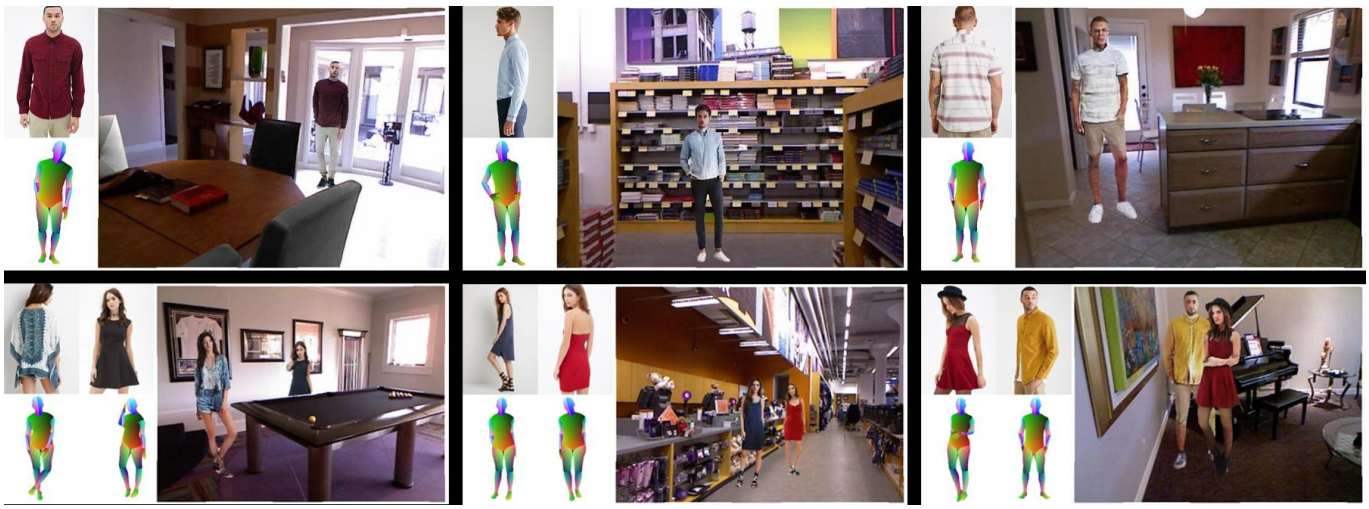


Fig T1.2.2 Sample images generated by our framework. For each example, we show the source image, the target 3d body mesh, and a scene with a geometrically plausible placement of the synthesized person. Please note that our framework allows for a positioning behind various objects, and the insertion of multiple people without breaking any geometrical scene properties.

Reference

M. Zanfir , E. Oneata , A. Popa, A. Zanfir , C. Sminchisescu, “Human Synthesis and Scene Compositing”, Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020)

Understanding 3d human interactions is fundamental for fine grained scene analysis and behavioural modeling. However, most of the existing models focus on analyzing a single person in isolation, and those who process several people focus largely on resolving multi-person data association, rather than inferring interactions. This may lead to incorrect, lifeless 3d estimates, that miss the subtle human contact aspects–the essence of the event–and are of little use for detailed behavioral understanding. This paper addresses such issues and makes several contributions: (1) we introduce models for interaction signature estimation (ISP) encompassing contact detection, segmentation, and 3d contact signature prediction; (2) we show how such components can be leveraged in order to produce augmented losses that ensure contact consistency during 3d reconstruction; (3) we construct several large datasets for learning and evaluating 3d contact prediction and reconstruction methods; specifically, we introduce CHI3D, a lab-based accurate 3d motion capture dataset with 631 sequences containing 2, 525 contact events, 728, 664 ground truth 3d poses, as well as FlickrCI3D, a dataset of 11, 216 images, with 14, 081 processed pairs of people, and 81, 233 facet-level surface correspondences within 138, 213 selected contact regions. Finally, (4) we present models and baselines to illustrate how contact estimation supports meaningful 3d reconstruction where essential interactions are captured. Models and data are made available for research purposes at http://vision.imar.ro/ci3d. To get a sense of the results of our method, please have a look at fig. T1.2.3 below.

Fig. T1.2.3 3D human pose and shape reconstructions using contact constraints of different granularity. The first column shows the RGB images followed by their reconstructions without contact information (column 2), using contacts based on 37 and 75 regions, respectively (columns 3 & 4), and using facet-based correspondences (column 5). While using facet-based constraints provides the most accurate estimates, reasonable results can be obtained even for coarser (region) assignments.

Reference

M. Fieraru, M. Zanfir, E. Oneata, A. Popa. V. Olaru, C. Sminchisescu, “Three-dimensional Reconstruction of Human Interactions”, Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

Monocular estimation of three dimensional human self-contact is fundamental for detailed scene analysis including body language understanding and behaviour modeling. Existing 3d reconstruction methods do not focus on body regions in self-contact and consequently recover configurations that are either far from each other or self-intersecting, when they should just touch. This leads to perceptually incorrect estimates and limits impact in those very fine-grained analysis domains where detailed 3d models are expected to play an important role. To address such challenges we detect self-contact and design 3d losses to explicitly enforce it. Specifically, we develop a model for Self-Contact Prediction (SCP), that estimates the body surface signature of selfcontact, leveraging the localization of self-contact in the image, during both training and inference. We collect two large datasets to support learning and evaluation: (1)

HumanSC3D, an accurate 3d motion capture repository containing 1, 032 sequences with 5, 058 contact events and 1, 246, 487 ground truth 3d poses synchronized with images collected from multiple views, and (2) FlickrSC3D, a repository of 3, 969 images, containing 25, 297 surface-to-surface correspondences with annotated image spatial support. We also illustrate how more expressive 3d reconstructions can be recovered under self-contact signature constraints and present monocular detection of face-touch as one of the multiple applications made possible by more accurate self-contact models. Examples of qualitative results of our method can be seen in fig. T1.2.4 below.

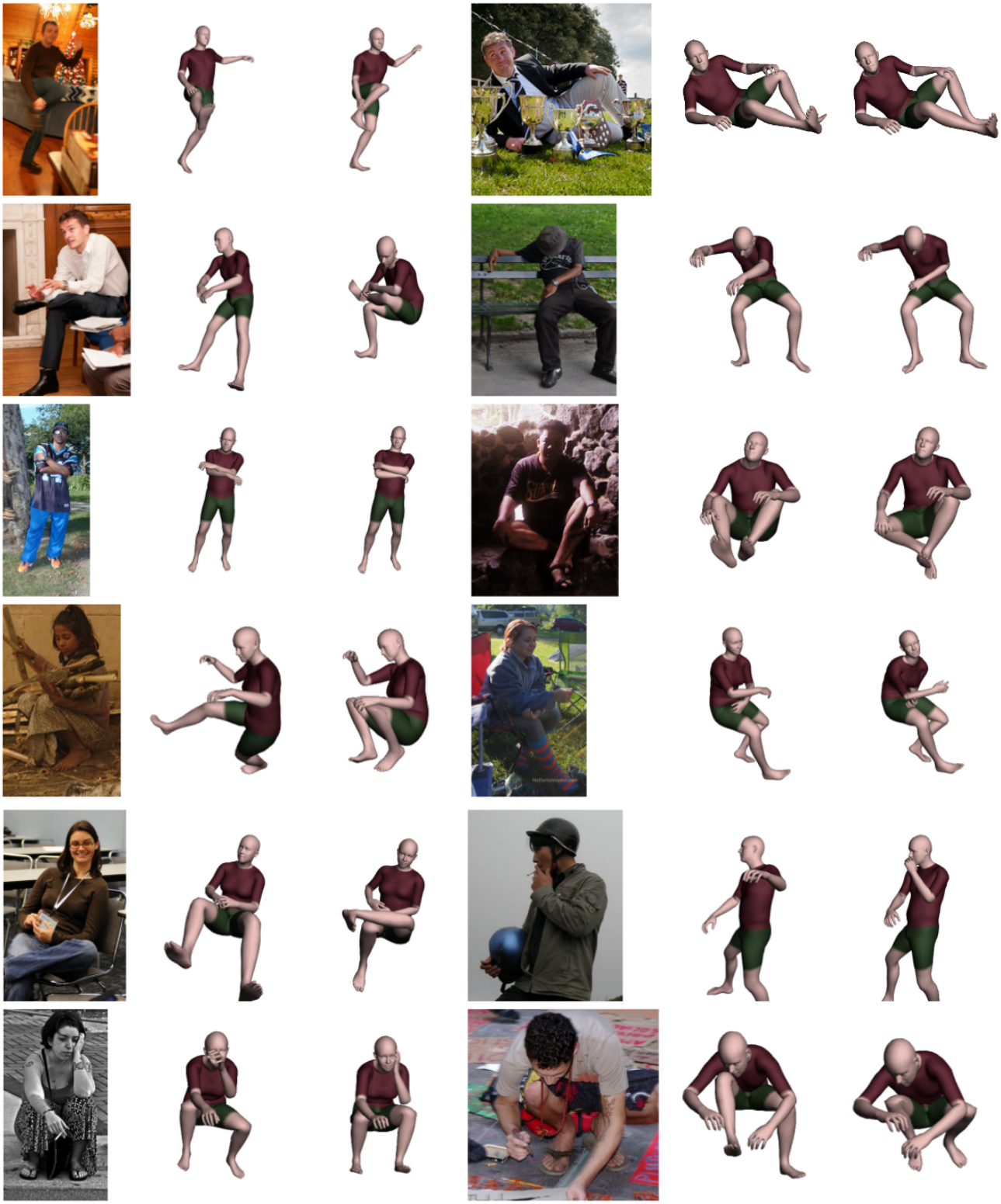


Fig. T1.2.4 3D pose and shape reconstructions using our annotated self-contact data. Left: Original image. Center: Reconstruction without considering the self-contact and the associated loss. Right: Reconstruction that uses the self-contact annotations and the corresponding loss.

Reference

M. Fieraru, M. Zanfir, E. Oneata, A. Popa, V. Olaru, C. Sminchisescu, “Learning Complex 3D Human Self-Contact”, accepted at the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)

In order to properly accommodate the aerial-based scenario, we then moved the focus towards monocular depth estimation (MDE) algorithms. Since the MDE problem is known to be ill-posed (an infinity of 3D scenes can be generated from a 2D image), in “Semi-Global Optimization for Classification-Based Monocular Depth Estimation” we firstly tried to constrain the MDE problem by applying geometric cues. To this end, we proposed a novel method that includes mathematical constraints into the MDE through a new stereo-based global optimization. Thus, we transformed the features extracted from the last layer of a MDE CNN into a stereo-like cost volume. This new volume is then optimized according to the semi-global matching (SGM) technique, which ensures that the depth map is globally consistent through the smoothness constraints.

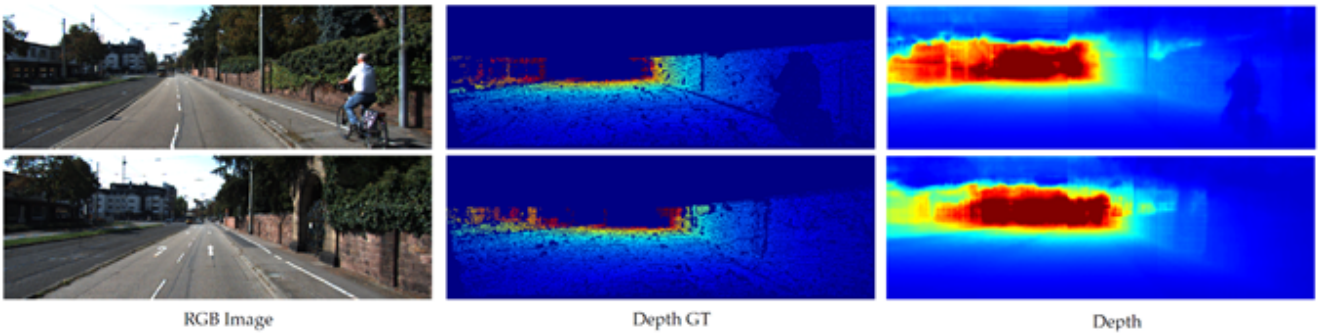


Fig. T1.2.7 Results from the paper “Semi-Global Optimization for Classification-Based Monocular Depth Estimation”. The depth map is accurately generated, by using the novel stereo-based global optimization. The results also present a higher confidence, due to the geometrical constraints introduced by the method

Reference

VC. Miclea, S. Nedevschi, “Semi-Global Optimization for Classification-Based Monocular Depth Estimation”, *Proceedings of 2020 IEEE International Conference on Intelligent Robots and Systems (IROS2020)*, Las Vegas, SUA, October 25-29, 2020,

Another problem inherent to camera-based depth perception systems (including MDE and stereo-based ones) is dealing with long-distance objects. In order to alleviate this issue, in “A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithm”, we proposed a novel unified method that captures relevant information from the MDE/stereo features and it uses it to learn a (sub-pixel) interpolation function such that wrongly estimated points in the far range are thoroughly corrected. The additional optimization constraints and the long-distance correction

methods prove that state of the art MDE methods can be further refined, generating depth maps that are more accurate, more reliable and robust.

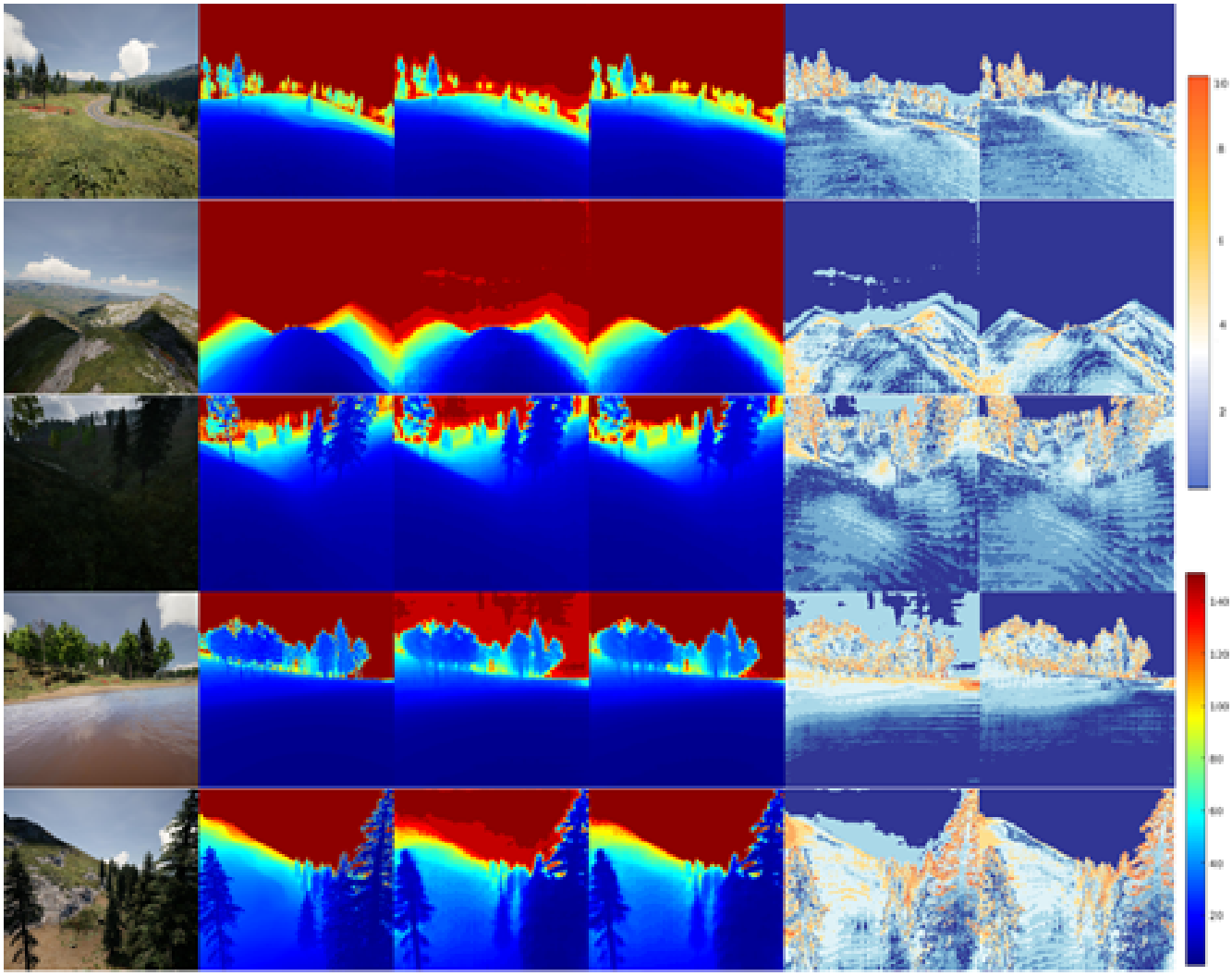


Fig. T1.2.8 Results from the article “A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithms”. MidAir images (synthetic) are used as input for the CNN, which produces highly accurate results (error images are presented on the last two columns)

Reference

VC. Miclea, S. Nedevschi , “A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithms”, *Proceedings of 2020 IEEE Intelligent Vehicles Symposium (IV2020)*, , Las Vegas, SUA.October, 19–November 13, 2020

Finally, since the end goal of this task is dealing with UAV-based perception, we tackled this problem as well. Thus, we introduced a novel MDE system, capable of working on complex aerial images, captured from a medium distance from a drone. The method proposes an original CNN, particularly adapted to such scenarios by introducing a novel feature extractor, a new scene understanding module and a new multi-task loss that combines state of the art MDE methods. An important part of this work was the

development of a novel fully-differentiable softmax transformation CNN layer that facilitates a better convergence for the network. The method can also benefit from the aforementioned refinement proposals, increasing the robustness by using the global optimization and dealing with objects at large distances. The proposed CNN proves to provide the most accurate results for depth generation from aerial images. Furthermore, it proves a high flexibility, being evaluated on images captured in a large variety of scenarios (form multiple positions, with different orientations and on multiple scenes).

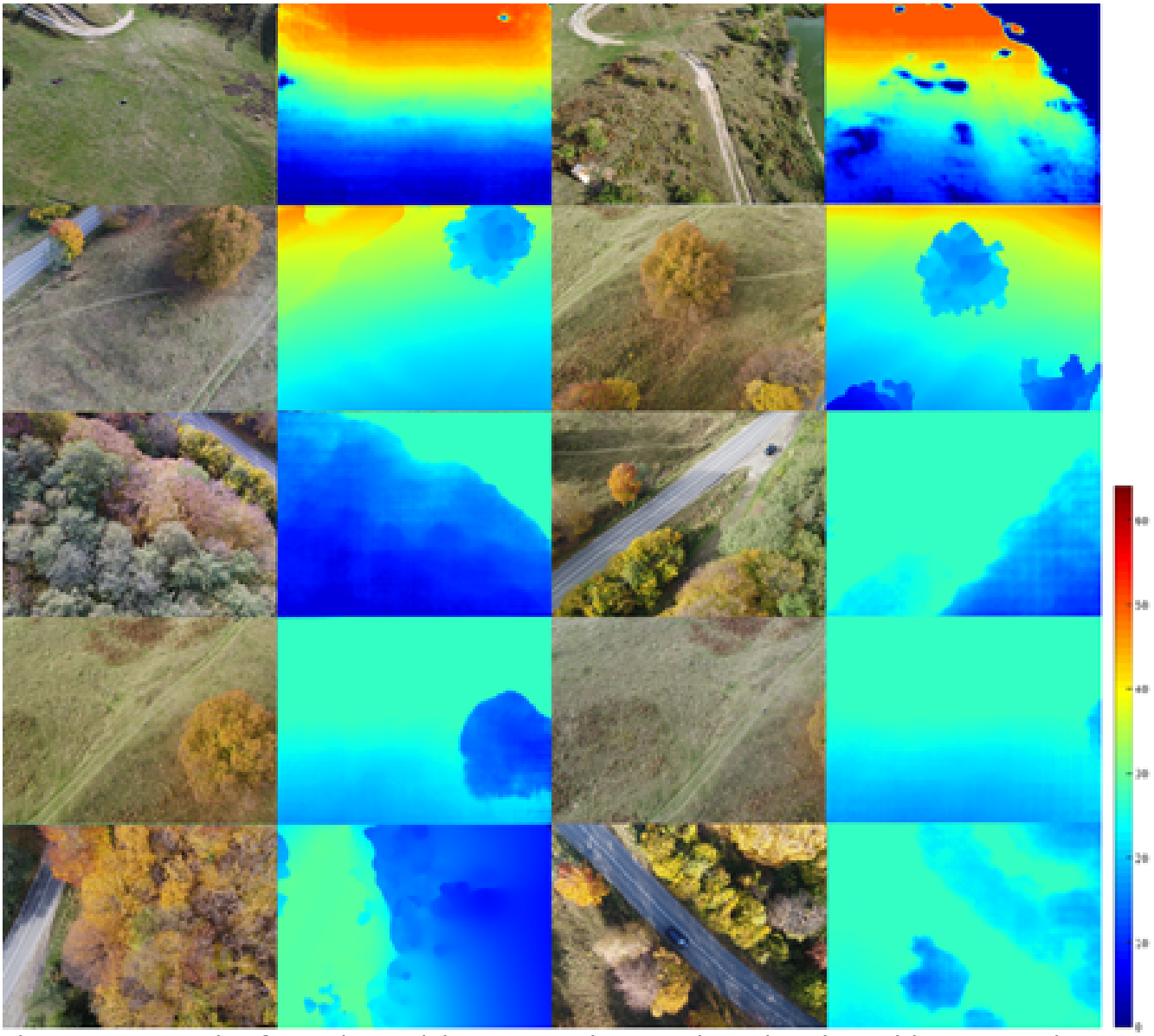


Fig. T1.2.9 Results from the article “Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception”. The method produces very accurate depth results on real-life images (Nadir and at various other angles), captured from a real drone, in various scenarios (fields, forests)

Reference

V. Miclea, S. Nedevschi, “Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception”, accepted at IEEE Transactions on Geoscience and Remote Sensing

A2. Visual Recognition and Localization

Task 2.1. Weakly-supervised semantic models with multiple components and partial responses

In “A Critical Evaluation of Aerial Data Datasets for Semantic Segmentation” we evaluated datasets recorded at various flight altitudes (DroneDeploy, Ruralscapes, and Mid-Air), in terms of class balance, training performance on the semantic segmentation task, and the ability to transfer knowledge from one set to another. Our findings showcase the strengths of the evaluated datasets, while also pointing out their shortcomings, and offering future development ideas and raising research questions. We believe that MidAir can be used for all learning tasks of our research problem, starting from object detection, semantic segmentation, to 3D reconstruction, localization, and mapping, since it contains ground truth annotation such as depth maps and semantic labels, but narrowing the semantic gap between real and synthetic data is a necessary task. AirSim is considered as a proper solution for developing frameworks that can solve the task of control, since it contains accurate drone physics modelling.

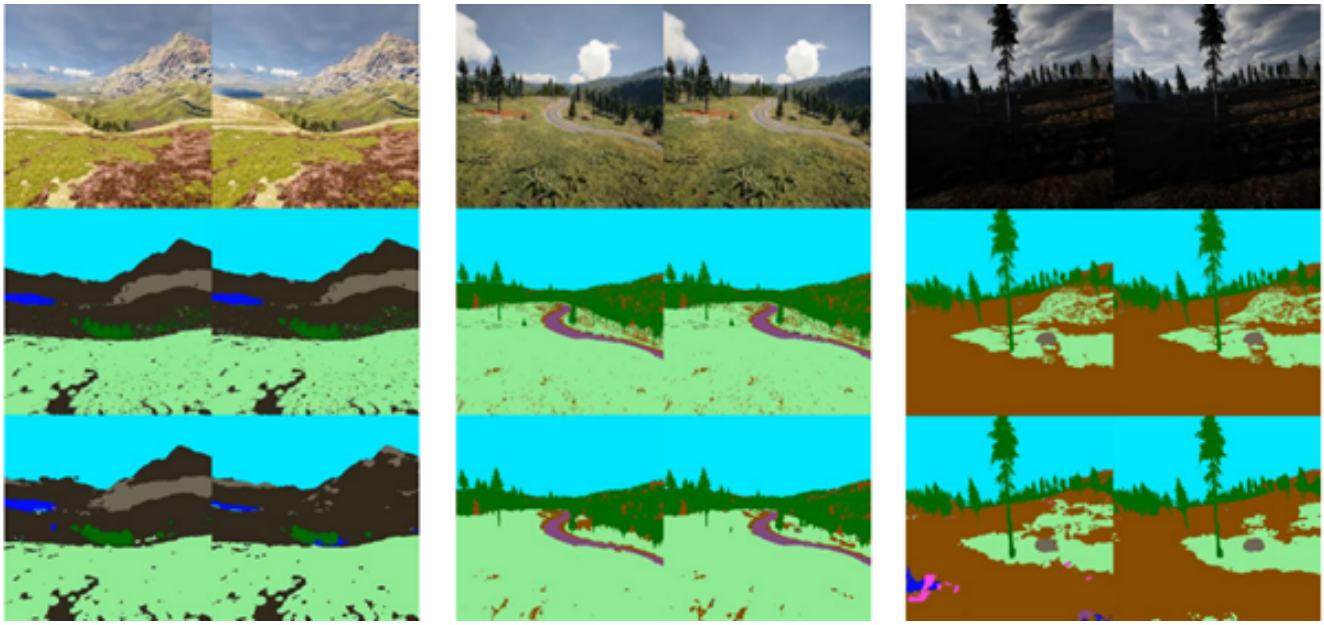


Fig. T2.1.3 Results from the paper ”A Critical Evaluation of Aerial Data Datasets for Semantic Segmentation”. Prediction results on the Mid-Air dataset, in 3 scenarios: mountain area, road in spring, and sunset in autumn. From top to bottom, the color image, the ground truth segmentation, and the semantic segmentation result. The first column is from MA50, while the second one – MA10.

Reference

B. C. Z. Blaga, S. Nedevschi, A Critical Evaluation of Aerial Datasets for Semantic Segmentation, *Proceedings of IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2020.

Image-to-image translation is an emerging method of computer vision dataset augmentation, which allows transferring the style of real life images onto synthetic ones,

making them more realistic. In our work, “Narrowing the semantic gap between real and synthetic data”, we propose an incremental improvement over the adversarial learning generator architectures used by image-to-image translation models. First, we managed to use a single network, instead of 2, thus creating a more memory efficient model, which allowed for end-to-end training on high resolutions. Second, inspired from recent work on semantic segmentation architectures, we enhanced our model by implying a multi-scale encoding and stylization phase, allowing for a better control over the contextual and spatial features. Given a synthetic image, our framework allows for its multimodal translation into the real domain. Our model shows promising results at narrowing the semantic gap between synthetic and real data.

Fig. T2.1.4 Results from the paper “Narrowing the semantic gap between real and synthetic data”. Sample images from the translation of GTA5 → Cityscapes. From left to right: the first image represents the original synthetic image, the second image represent the style which was applied, the third column show the image reconstruction using the given style and the last column images reconstructed with a random style sampled from the normal N (0, 1) distribution are displayed.

Reference

R. Beche, S. Nedevschi, “Narrowing the semantic gap between real and synthetic data”, *Proceedings of IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2020,

In ”Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences” a solution was developed for weakly supervised learning of aerial video semantic segmentation leveraging the relation between neighbouring frames. The system is composed of a static semantic segmentation, an optical flow and a linking network, which are chosen from existing architectures based on their high accuracy and low computational needs.

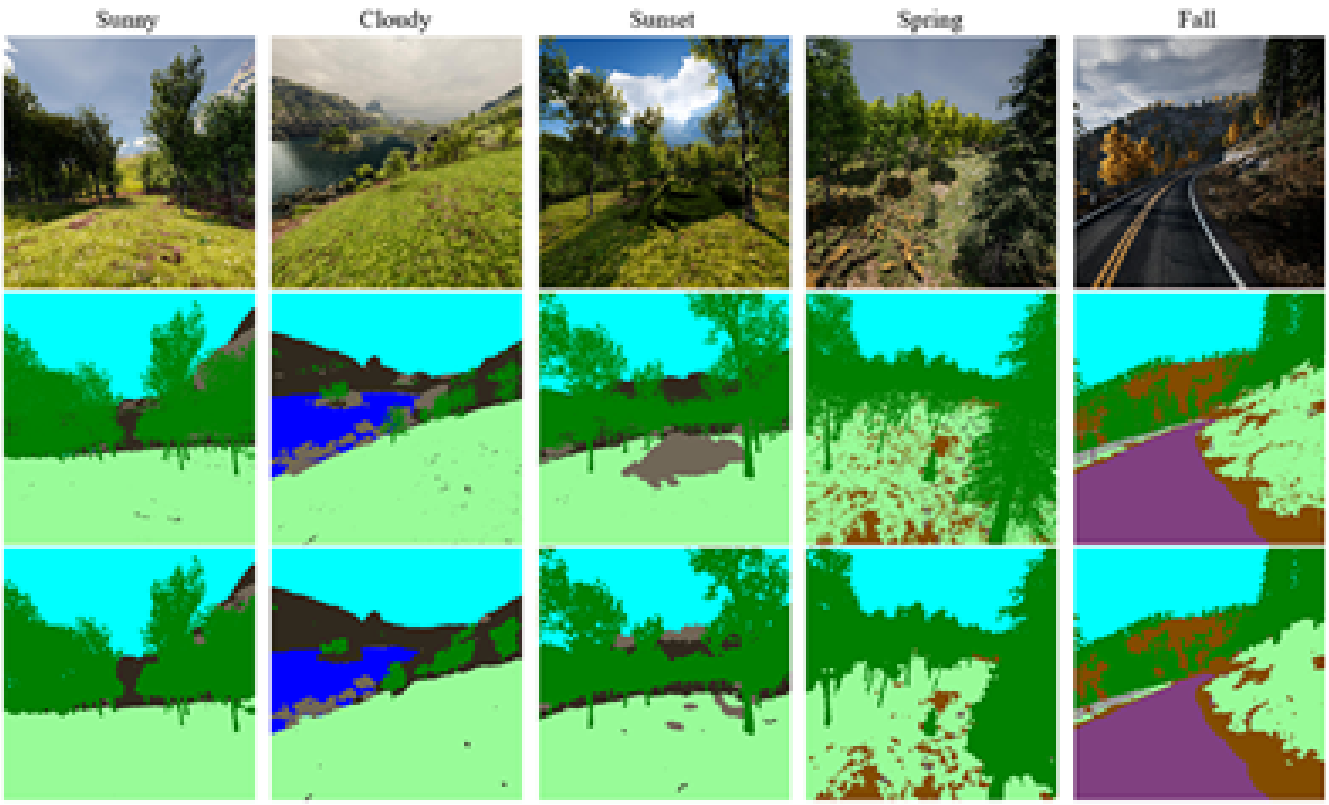


Fig. T2.1.5 Results of the paper “Weakly Supervised Semantic Segmentation Learning On UAV Video Sequences”. Results of the framework on the test set, presented for the 5 different scenarios from Mid-Air.

Reference

B. C. Z. Blaga and S. Nedevschi**,** ”Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences”, submitted to European Signal Processing Conference, 2021

Z. Blaga, Aerial Dataset for Semantic Segmentation, to be published

Task 2.2. Active and adversarial learning structures and methods for dynamic data.

In “Real-Time Panoptic Segmentation with Prototype Masks for Automated Driving”, we design a one-stage network for panoptic segmentation that is lightweight, accurate and much faster than the previous two-stage solutions. Our network learns semantic masks but does not directly learn instance masks. In order to obtain instance logits, our network learns a fixed number of scene prototype masks, which are assembled guided by a proposal-based weighting scheme. We propose a recalibration scheme for panoptic logits refinement. Our solution is the fastest on the Cityscapes benchmark and achieves comparable results with other state-of-the-art methods.

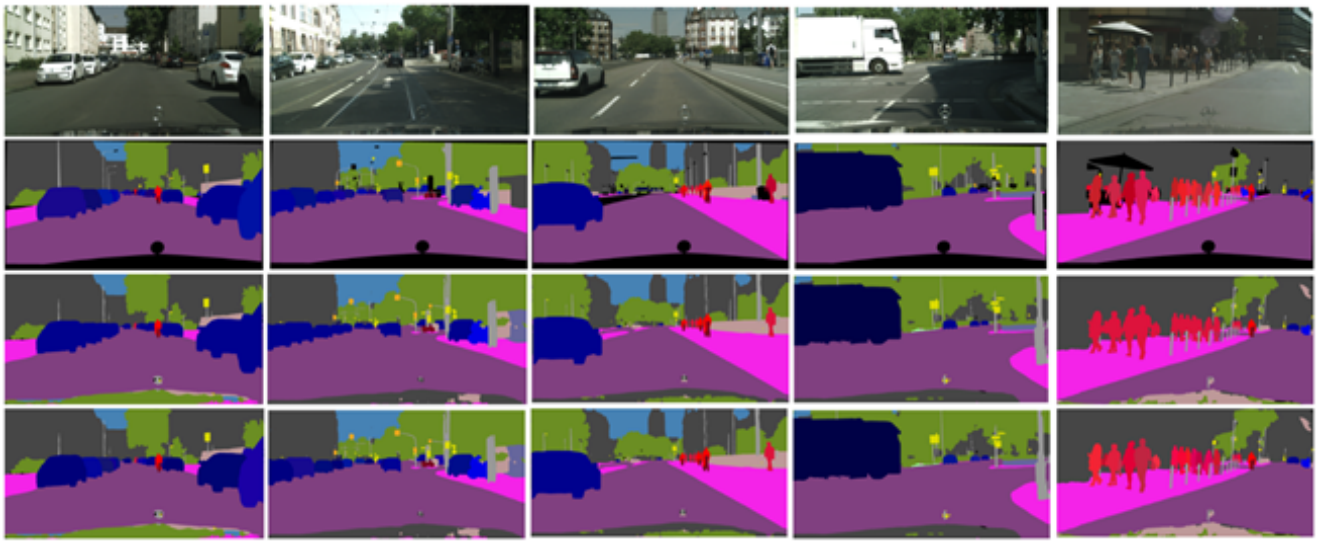


Fig. T2.2.4 Results from the paper “Real-Time Panoptic Segmentation with Prototype Masks for Automated Driving”. From top to bottom: image, panoptic ground truth, semantic segmentation, panoptic segmentation. In the panoptic segmentation the color encodes the class and the instance identifier. Our network can correctly segment large and small scale objects and also occluded objects.

Reference

A. Petrovai, S. Nedevschi, “Real-Time Panoptic Segmentation with Prototype Masks for Automated Driving”, Proceedings of 2020 IEEE Intelligent Vehicles Symposium (IV2020), Las Vegas, SUA, October 19–November 13, 2020.

We reach state-of-the-art accuracy on the Cityscapes dataset with the fast and accurate network proposed in “SAPSNet: A Soft Attention Panoptic Segmentation Network”. In this work, we introduce a fast and accurate single-stage panoptic segmentation network that employs a shared feature extraction backbone and dual-decoders that learns semantic and instance-level attention masks. Guided by object proposals, our new instance-level decoder learns instance specific soft attention masks based on spatial embeddings, pixel offsets to the object center. The panoptic output incorporates semantic masks for background classes, while the foreground classes are attended by the soft instance masks. Training and inference processes are unified and no post-processing operations are necessary. Our model outperforms state-of-the-art approaches that aim real-time performance in both inference speed and quality and achieves competitive results on the Cityscapes dataset.

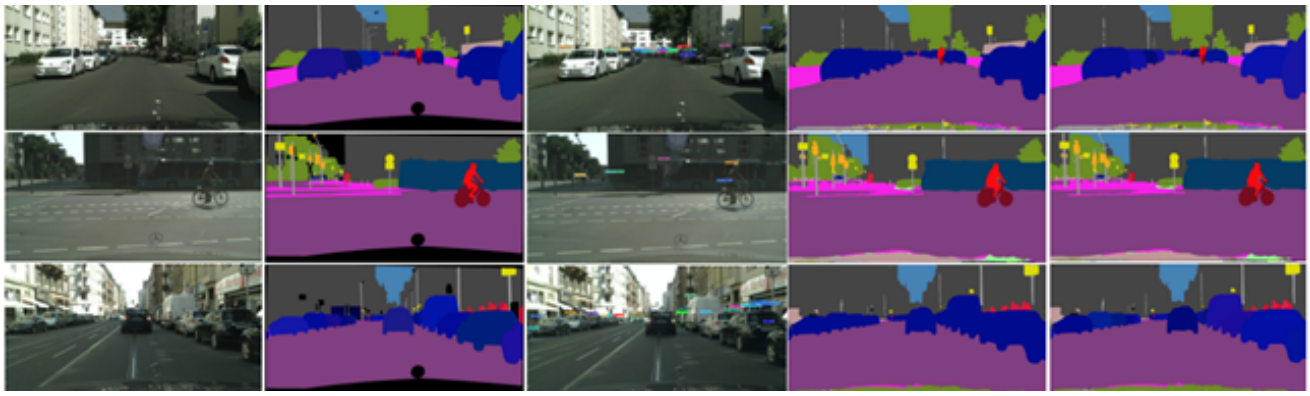


Fig. T2.2.5 Results from the paper “SAPSNet: A Soft Attention Panoptic Segmentation”. From left to right: image, panoptic segmentation ground truth, object detection, semantic

segmentation and panoptic segmentation. Our network can accurately segment object of various sizes and can handle difficult scenarios with occlusions.

Reference

A. Petrovai, S. Nedevschi, “SAPSNet: A Soft Attention Panoptic Segmentation”, submitted to IEEE Transactions on Image Processing

In the paper “Video Semantic Segmentation leveraging Dense Optical Flow” we have also studied recurrent layers that are able to temporally propagate semantic information by means of optical flow. The gated propagation modules and optical flow are jointly trained for increased performance. We employ a very fast segmentation network to provide image level segmentation. The semantic output from previous frames is spatially aligned with the current frame with the dense optical flow network. A conv-GRU module fuses the current and aligned segmentation. Experiments on synthetic and real datasets, Virtual KITTI and Cityscapes, show that temporal information is important for increased performance accuracy on video sequences.

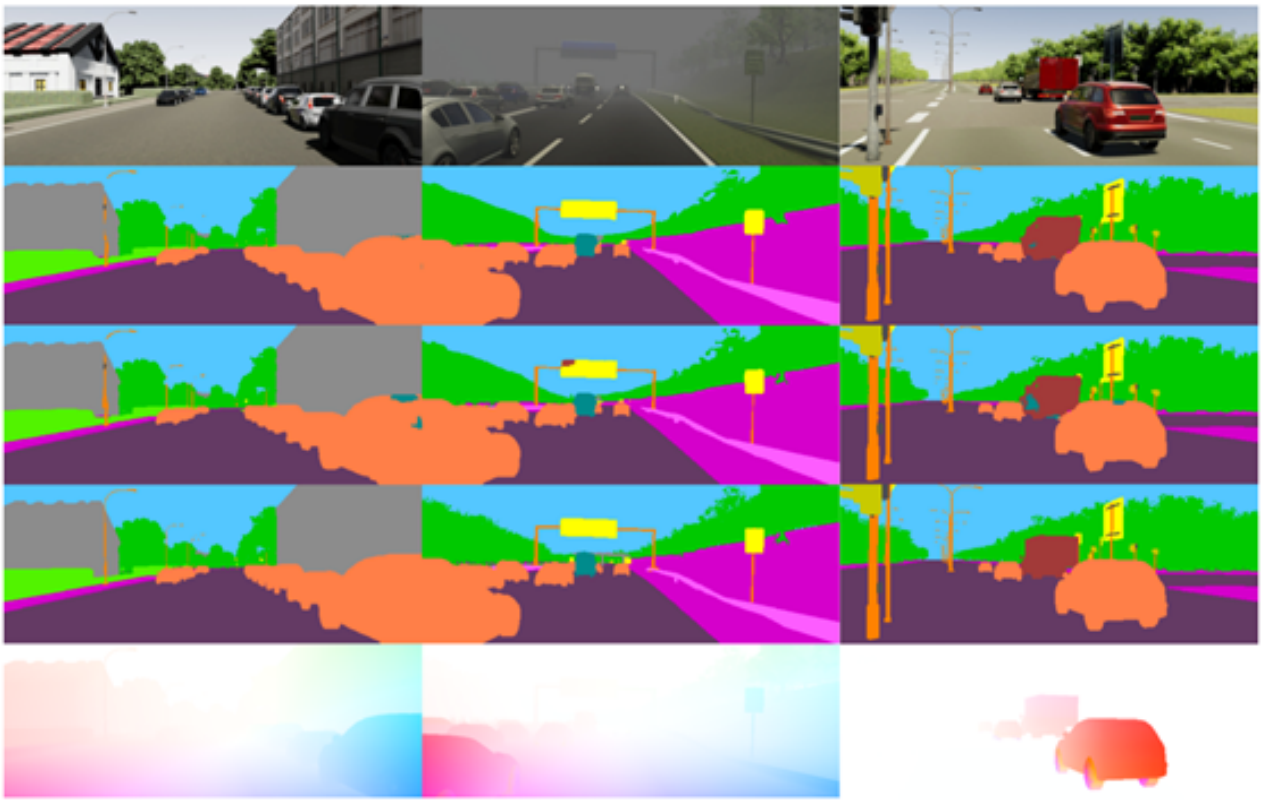


Fig. T2.2.6 Results from the paper “Video Semantic Segmentation leveraging Dense Optical Flow”. Examples of segmentation on Virtual KITTI test subset. From top to bottom are the current frame, refined segmentation, static segmentation, ground truth and backward optical flow.

Reference

V. Lup, S. Nedevschi, “Video Semantic Segmentation leveraging Dense Optical Flow”, *Proceedings of 2020 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP2020)*, September 3-5, 2020, Cluj-Napoca, Romania.

We developed a 360-degree perception system that has been integrated in a prototype vehicle "**Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving**". We implemented deep learning based semantic virtual cameras that provide semantic, instance and panoptic segmentation by processing images from five cameras: four fisheye cameras and one narrow field-of-view camera. Fisheye cameras provide near-range 360-degree coverage, while the 60-degree front camera extends the detection range three time. We meet requirements of high accuracy and low processing time in order to enable fully automated navigation of the vehicle. We create a large scale dataset of fisheye and perspective image with semantic and instance annotations, that has been used for training the networks. The automated vehicle equipped with our 2D perception system has been successfully demonstrated in urban areas after extensive experiments.

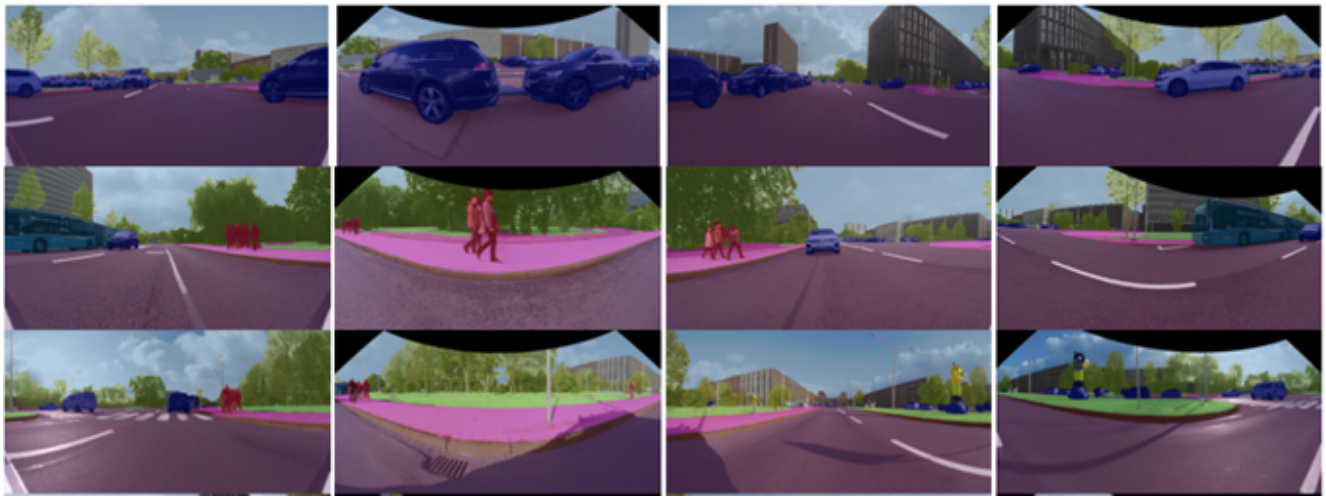


Fig. T2.2.7 Results from the paper "**Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving**". Semantic segmentation of unwarped fisheye images. We process four images from the fisheye 160◦ horizontal field-of-view cameras which provide 360◦ coverage around the vehicle. Each camera views a different direction around the vehicle: front, right, rear and left.

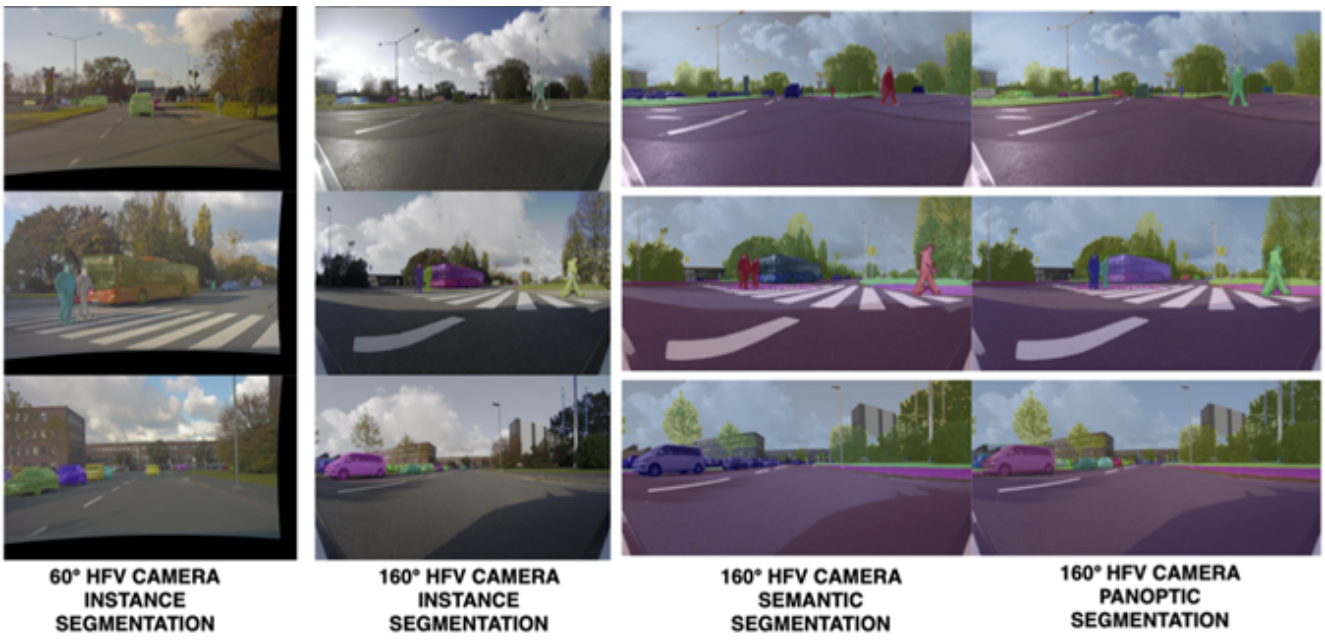


Fig. T2.2.8 Results from the paper "**Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving**". The front area of the vehicle is

covered by two cameras: a narrow 60◦ horizontal field-of-view camera which provides instance segmentation at increased depth and a wider 160◦ horizontal field-of-view camera, which provides instance, semantic and panoptic segmentation for the near-range.

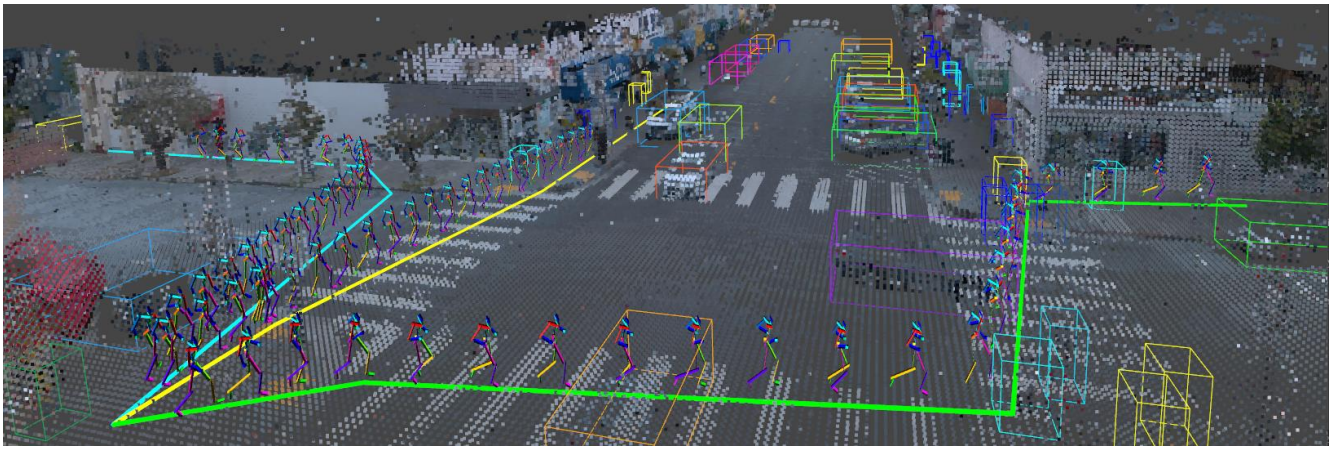
Reference

A. Petrovai, S. Nedevschi, “Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving”, submitted to *IEEE Transactions on Intelligent Transportation Systems*

A3. Semantic Optimal Control

Task 3.1. Direct and Inverse Optimal control

We present a model for generating 3d articulated pedestrian locomotion in urban scenarios, with synthesis capabilities informed by the 3d scene semantics and geometry. We reformulate pedestrian trajectory forecasting as a structured reinforcement learning (RL) problem. This allows us to naturally combine prior knowledge on collision avoidance, 3d human motion capture and the motion of pedestrians as observed e.g. in Cityscapes, Waymo or simulation environments like Carla. Our proposed RL-based model allows pedestrians to accelerate and slow down to avoid imminent danger (e.g. cars), while obeying human dynamics learnt from in-lab motion capture datasets. Specifically, we propose a hierarchical model consisting of a semantic trajectory policy network that provides a distribution over possible movements, and a human locomotion network that generates 3d human poses in each step. The RL-formulation allows the model to learn even from states that are seldom exhibited in the dataset, utilizing all of the available prior and scene information. Extensive evaluations using both real and simulated data illustrate that the proposed model is on par with recent models such as S-GAN, ST-GAT and S-STGCNN in pedestrian forecasting, while outperforming these in collision avoidance. We also show that our model can be used to plan goal reaching trajectories in urban scenes with dynamic actors. Fig. T3.1.1 and T3.1.2 show qualitative results .



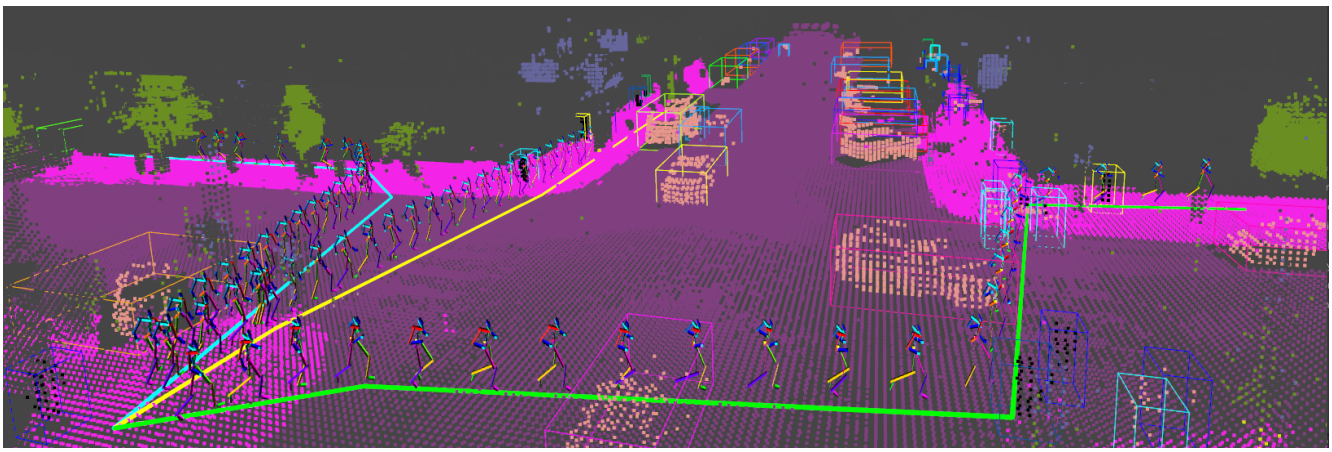


Fig T3.1.1 Pedestrian trajectories and poses generated by our agent on a Waymo scene. RGB and semantic point clouds of the scene are shown in the top and bottom images, respectively. A local neighborhood of these point clouds are observed by the agent. Coloured lines on the ground show different trajectories taken by the agent when initialized with varying agent histories. The agent crosses the roads without collisions. Cars and other pedestrians in the scene are shown as positioned in the first frame and are surrounded by bounding boxes for clarity.

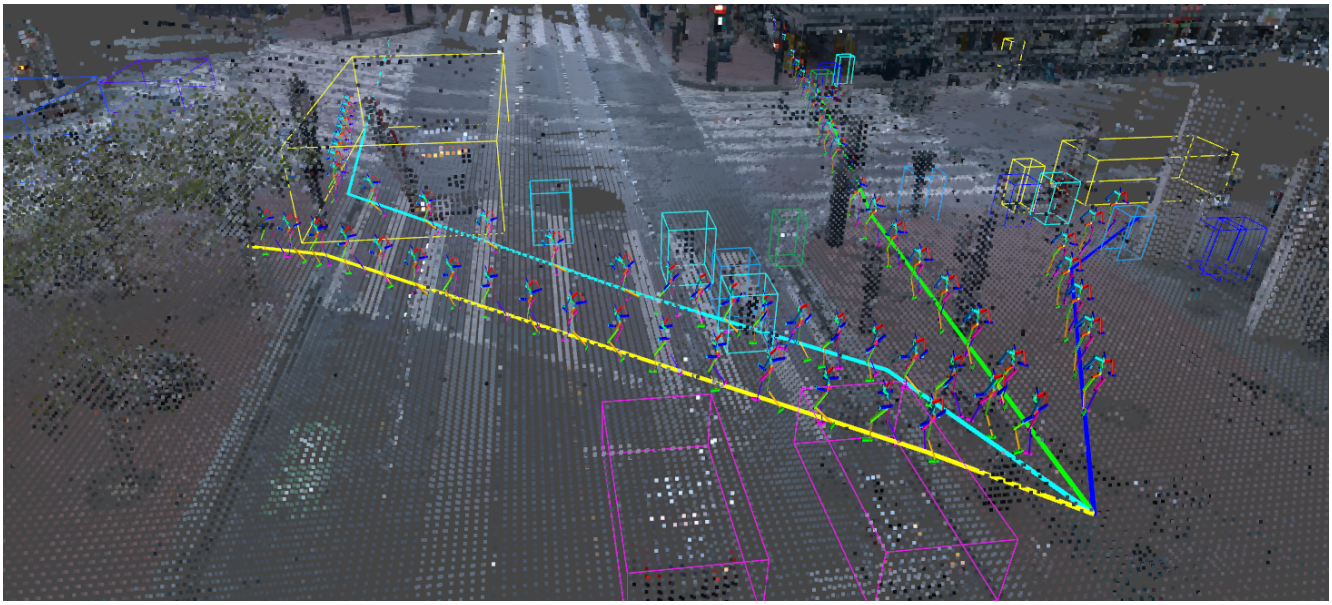


Fig. T3.1.2 SPL agent trajectories on the Waymo dataset, showing the pedestrian taking a number of different paths depending on how the agent history is initialized. Cars and other pedestrians are indicated with 3d bounding boxes. pedestrian trajectories. It should be noted that the collision-aware SPL agent travels slower than BC to avoid collisions, which results in shorter trajectories on average. However SPL's trajectories are three times longer than S-STG(CNN) with half of the collisions. The SPL model has the second lowest ADE after BC (which shares SPL's architecture) on the Waymo dataset. The SPL model is the only model to perform well on trajectory forecasting on both simulated and real data, while outperforming all models in collision avoidance. Qualitative examples of the SPL agent (without goals) are shown in Fig. T3.1.1, Fig. T3.1.2

Reference

M. Priisalu, C. Paduraru, A. Pirinen, C. Sminchisescu, “Semantic Synthesis of Pedestrian Locomotion”, Proceedings of the Asian Conference on Computer Vision (ACCV), 2020

Most 3d human pose estimation methods assume that input – be it images of a scene collected from one or several viewpoints, or from a video – is given. Consequently, they focus on estimates leveraging prior knowledge and measurement by fusing information spatially and/or temporally, whenever available. In this paper we address the problem of an active observer with freedom to move and explore the scene spatially – in ‘time-freeze’ mode – and/or temporally, by selecting informative viewpoints that improve its estimation accuracy. Towards this end, we introduce Pose-DRL, a fully trainable deep reinforcement learning-based active pose estimation architecture which learns to select appropriate views, in space and time, to feed an underlying monocular pose estimator. We evaluate our model using single- and multi-target estimators with strong results in both settings. Our system further learns automatic stopping conditions in time and transition functions to the next temporal processing step in videos. In extensive experiments with the Panoptic multi-view setup, and for complex scenes containing multiple people, we show that our model learns to select viewpoints that yield significantly more accurate pose estimates compared to strong multi-view baselines. Results of our method are qualitatively presented in fig T3.1.3 and T3.1.4.

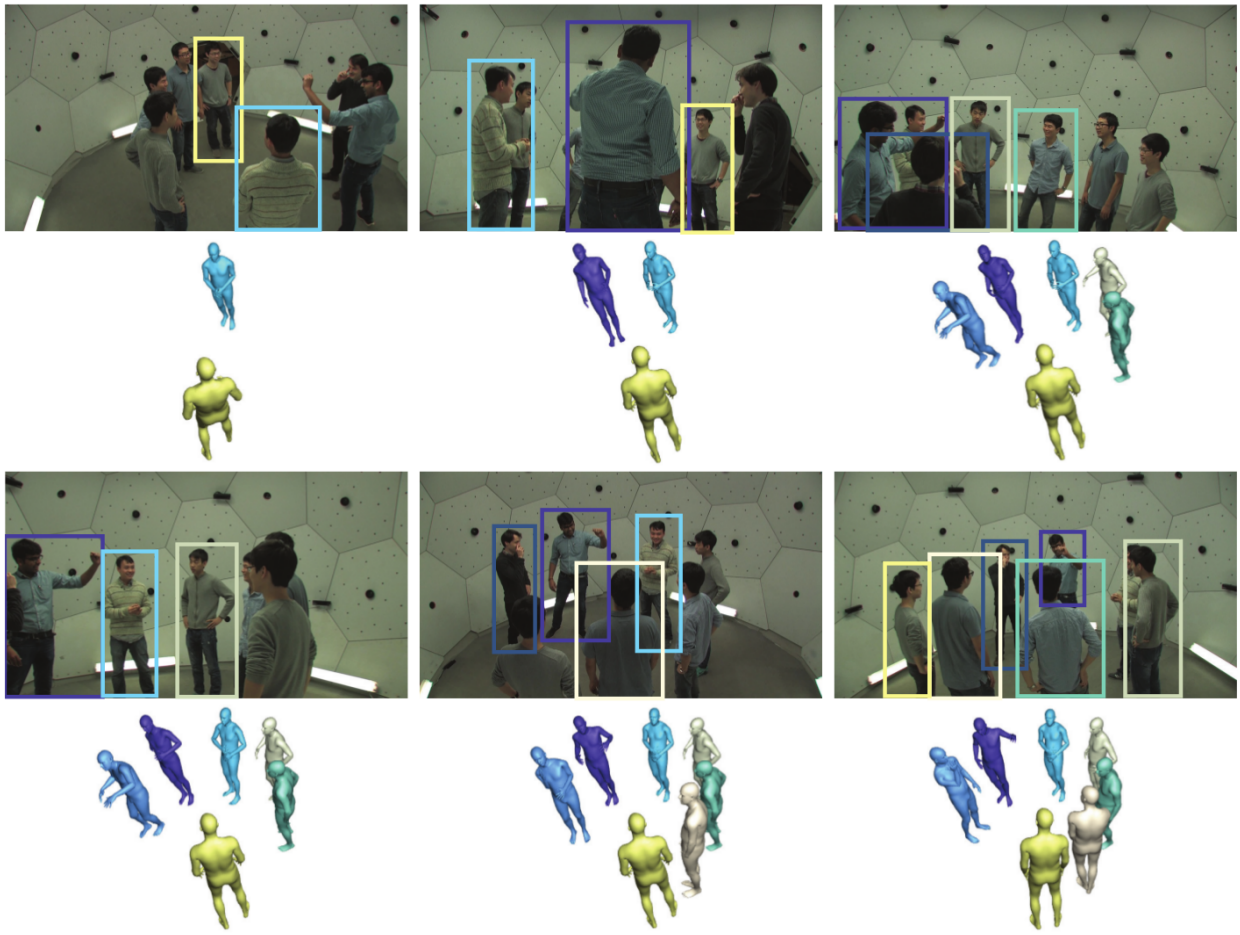


Fig. T3.1.3 Visualization of how Pose-DRL performs multi-target pose estimation for an Ultimatum test scene. In this example the agent sees six viewpoints prior to automatically continuing to the next active-view. The mean error decreases from 358.9

to 114.6 mm/joint. Only two people are detected in the initial viewpoint, but the number of people detected increases as the agent inspects more views. Also, the estimates of already detected people improve as they get fused from multiple viewpoints.

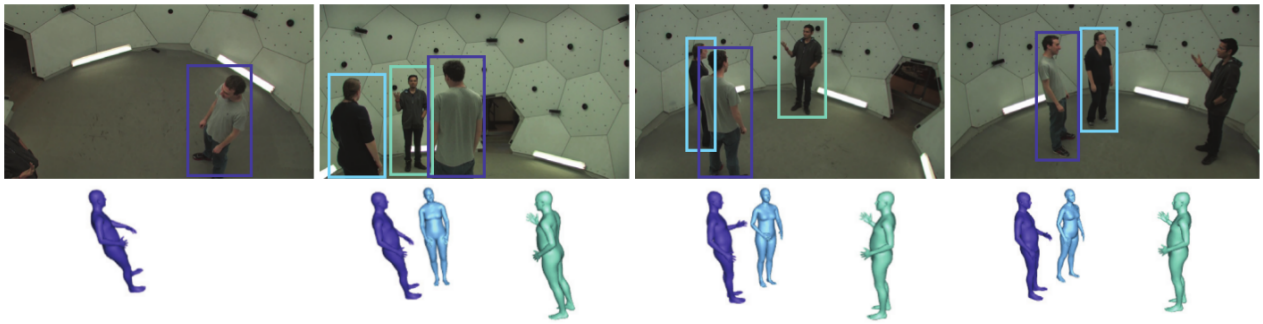


Fig. T3.1.4 Visualization of how Pose-DRL performs multi-target pose estimation for an Ultimatum validation scene. The agent chooses four viewpoints prior to automatically continuing to the next active-view. The mean error decreases from 334.8 to 100.9 mm/joint. Only one of the persons is visible in the initial viewpoint, and from a poor angle. This produces the first, incorrectly tilted pose estimate, but the estimate improves as the agent inspects more viewpoints. The two remaining people are successfully reconstructed in subsequent viewpoints.

Reference

E. Gärtner, A. Pirinen, C. Sminchisescu. “Deep Reinforcement Learning for Active Human Pose Estimation”, Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020

Task 3.2. Representations and Methods for Efficient Computation

We study the task of embodied visual active learning, where an agent is set to explore a 3d environment with the goal to acquire visual scene understanding by actively selecting views for which to request annotation. While accurate on some benchmarks, today's deep visual recognition pipelines tend to not generalize well in certain real-world scenarios, or for unusual viewpoints. Robotic perception, in turn, requires the capability to refine the recognition capabilities for the conditions where the mobile system operates, including cluttered indoor environments or poor illumination. This motivates the proposed task, where an agent is placed in a novel environment with the objective of improving its visual recognition capability. To study embodied visual active learning, we develop a battery of agents - both learnt and pre-specified - and with different levels of knowledge of the environment. The agents are equipped with a semantic segmentation network and seek to acquire informative views, move and explore in order to propagate annotations in the neighbourhood of those views, then refine the underlying segmentation network by online retraining. The trainable method uses deep reinforcement learning with a reward function that balances two competing objectives: task performance, represented as visual recognition accuracy, which requires exploring the environment, and the necessary amount of annotated data requested during active exploration. We extensively evaluate the proposed models using the photorealistic Matterport3D simulator and show that a fully learnt method outperforms comparable pre-specified counterparts, even when requesting fewer annotations. Qualitative results are shown in fig. T3.2.1 and T3.2.2.

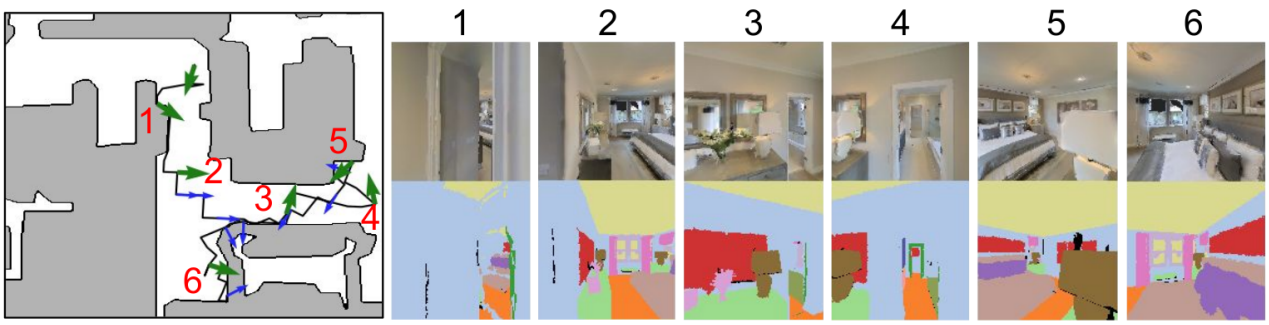


Fig T3.2.1 The first six requested annotations by the RL-agent in a room from the test set. Left: Map showing the agent's trajectory and the six first requested annotations (green arrows). The initially given annotation is not indicated with a number. Blue arrows indicate Collect actions. Right: For each annotation (numbered 1 - 6) the figures show the image seen by the agent and the ground truth received when the agent requested annotations. As can be seen, the agent quickly explores the room and requests annotations containing diverse semantic classes.

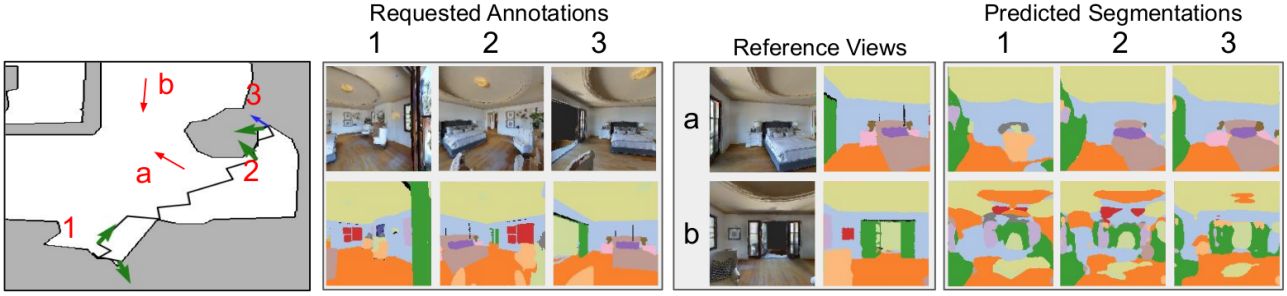


Fig T3.2.2 Example of the RL-agent's viewpoint selection and how its perception improves over time. We show results of two reference views after the first three annotations of the RL-agent. Left: Agent's movement path is drawn in black on the map. The annotations (green arrows) are numbered 1 - 3, and the associated views are shown immediately right of the map (the initially given annotation is not shown). Red arrows labeled a - b indicate the reference views. Right: Reference views and ground truth masks, followed by predicted segmentation after one, two and three annotations. Notice clear segmentation improvements as the agent requests more annotations. Specifically, note how reference view a improves drastically with annotation 2 as the bed is visible in that view, and with annotation 3 where the drawer is seen. Also note how segmentation improves for reference view b after the door is seen in annotation 3.

Reference

D. Nilsson , A. Pirinen, E. Gärtner , C. Sminchisescu. "Embodied Visual Active Learning for Semantic Segmentation", Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021)

A4. Systems Optimization and Integration

Task 4.1: Simultaneous localization and semantic mapping

For supporting the supervised learning tasks as well as for enabling evaluation of the supervised and self-supervised tasks, we have created an aerial imaging dataset using the DJI Matrice 210 V2 RTK drone [27]. It comprises over 1800 high-resolution images alongside video sequences from multiple flights over forest and open terrains, for which accurate positioning data is available. Using an open-source, aerial mapping software [28] based on traditional structure from motion and multi-view-stereo techniques [29] [30], textured 3D meshes were generated for each flight area, along with a 3D point cloud and a digital elevation model (at resolutions in the range of 5-10 cm/pixel). The 3D surface of the mesh enables us to generate, by reprojection, the dense depth image for each acquired color image. These pairs of color and depth images, corresponding to accurately 6D positioned camera poses, can then be used as the ground truth information for learning and evaluation processes. The intermediary area maps, which are accurately positioned based on DGPS information, also enable further visual localization tasks. The dataset is planned to be made publicly available.

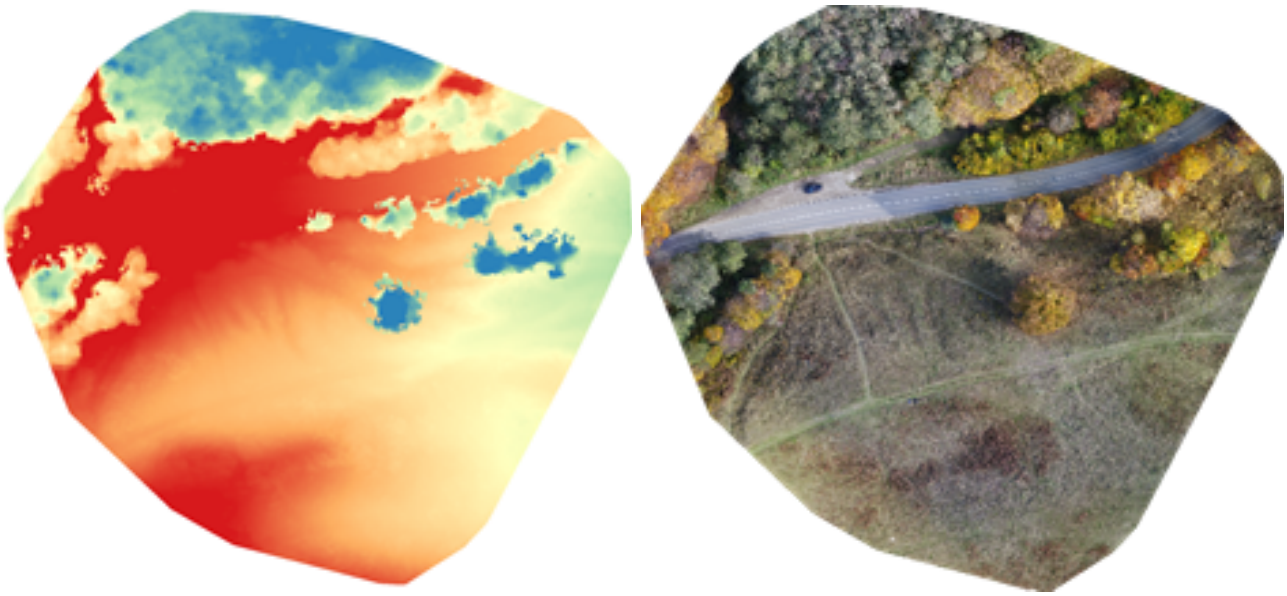


Fig. T4.1.1 Digital Surface Model (DSM) coloured based on elevation and Ortophoto resulting from the mapping process for one of the areas in the dataset

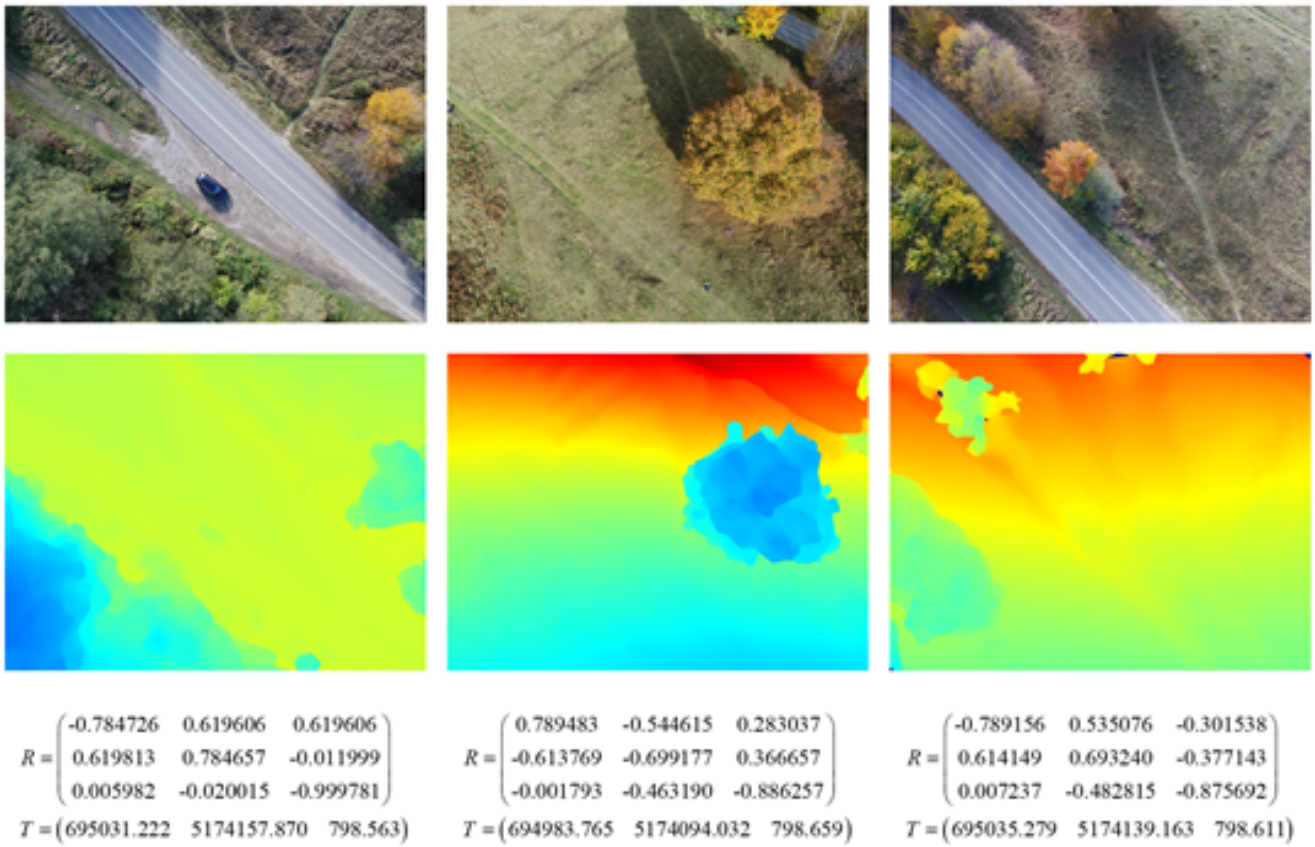
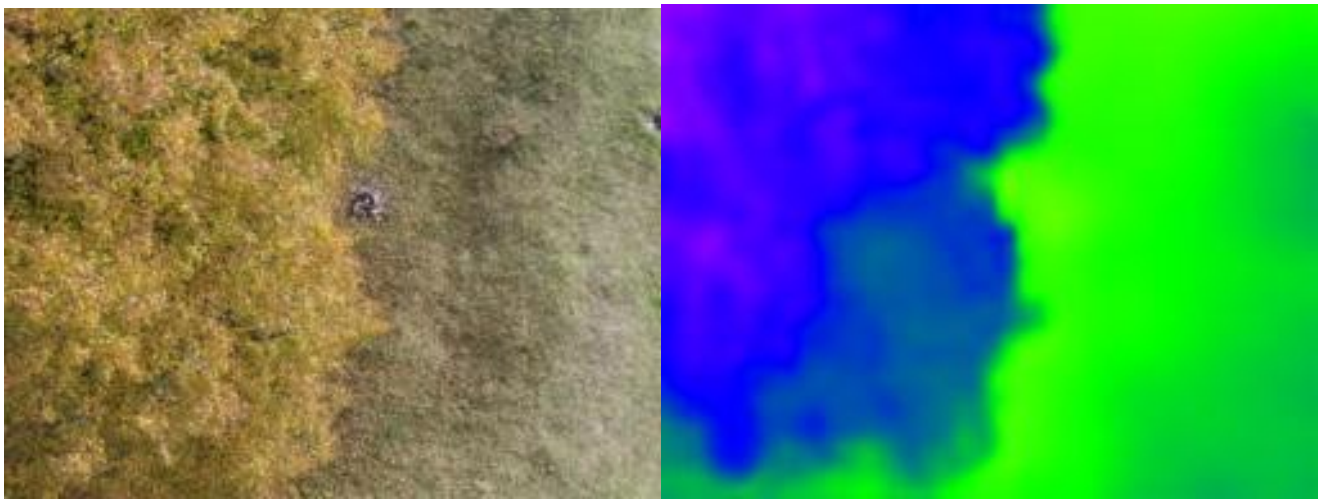


Fig. T4.1.2. Examples of the dense depth maps (second row) obtained by reprojecting the 3D mesh onto the original images (first row) for which the precise pose is available (third row)

A self-supervised depth and ego-motion estimation solution was experimented on the acquired video sequences and is under evaluation.



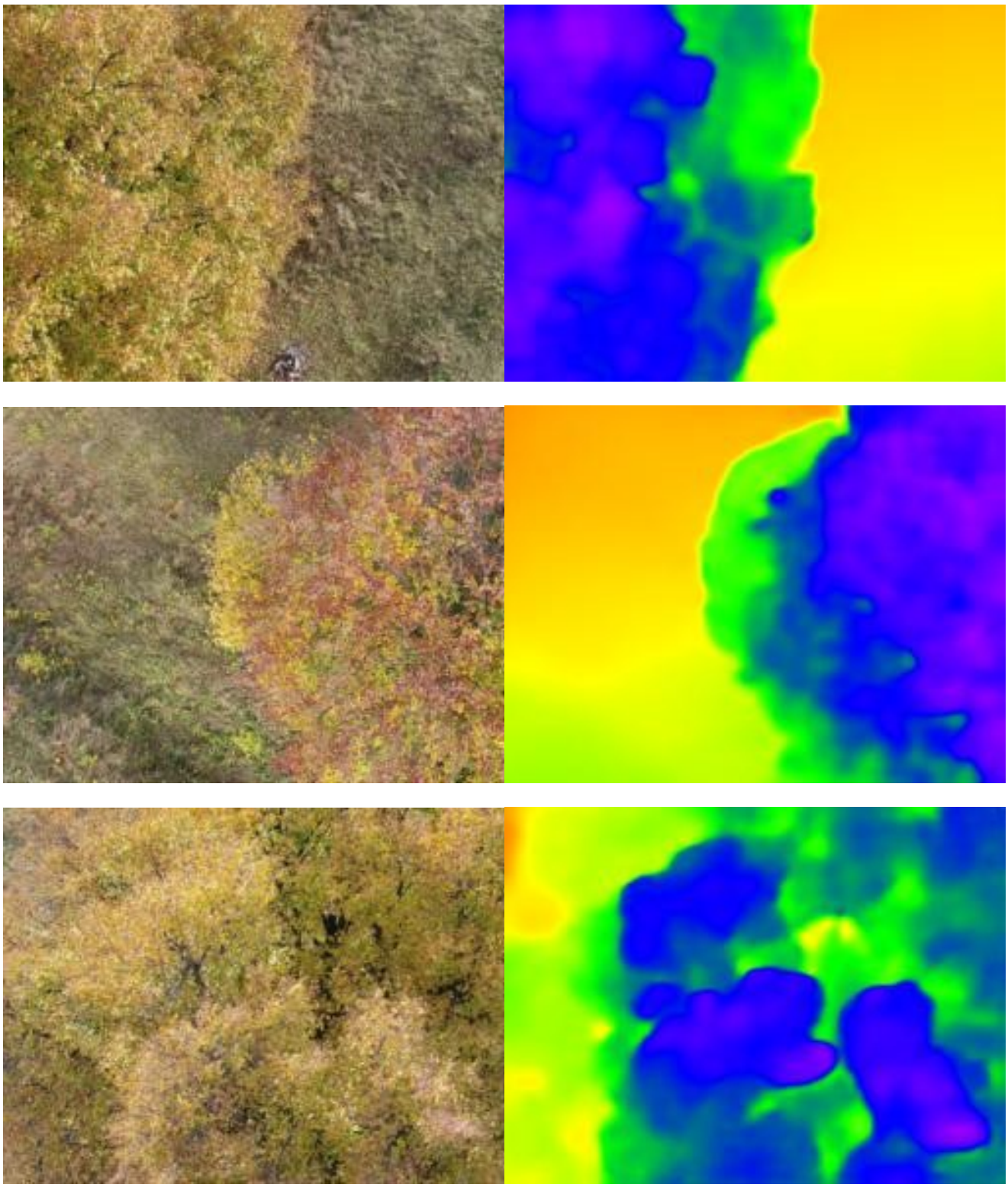


Fig. T4.1.3 Results of a *self-supervised depth and ego-motion estimation solution adapted for aerial images.*

Reference

H. Florea, Aerial Dataset for MDE and Mapping, to be published

Task 4.2: Integration of visual navigation and scene understanding.

In the paper "**Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion**" we present a real-time, 360-degree enhanced perception system which was successfully integrated onto an autonomous vehicle. The system is based on low-level fusion between 3D point clouds obtained from multiple LiDARs and semantic scene information obtained from multiple RGB cameras. The semantic, instance and panoptic segmentations of 2D data were computed using efficient and optimized deep-learning based algorithms, while the aligned 3D point clouds are segmented using a fast, traditional voxel-based solution. On top of the fused geometric and semantic data more effective detection, classification and localization algorithms were implemented.

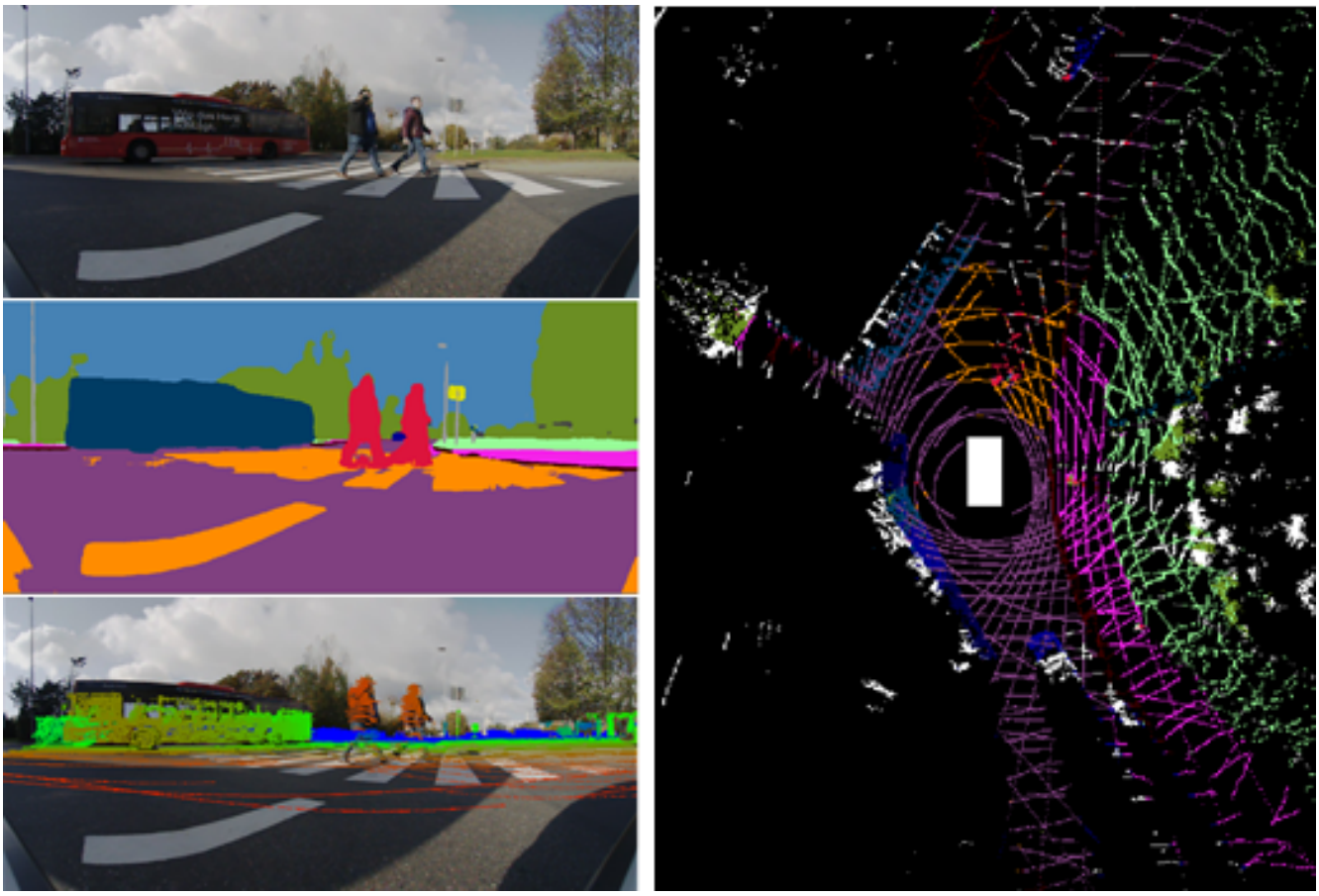


Fig. T4.1.4 Results from the paper "**Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion**". Low level fusion concept and Spatio-Temporal and Appearance Based Representation (STAR)

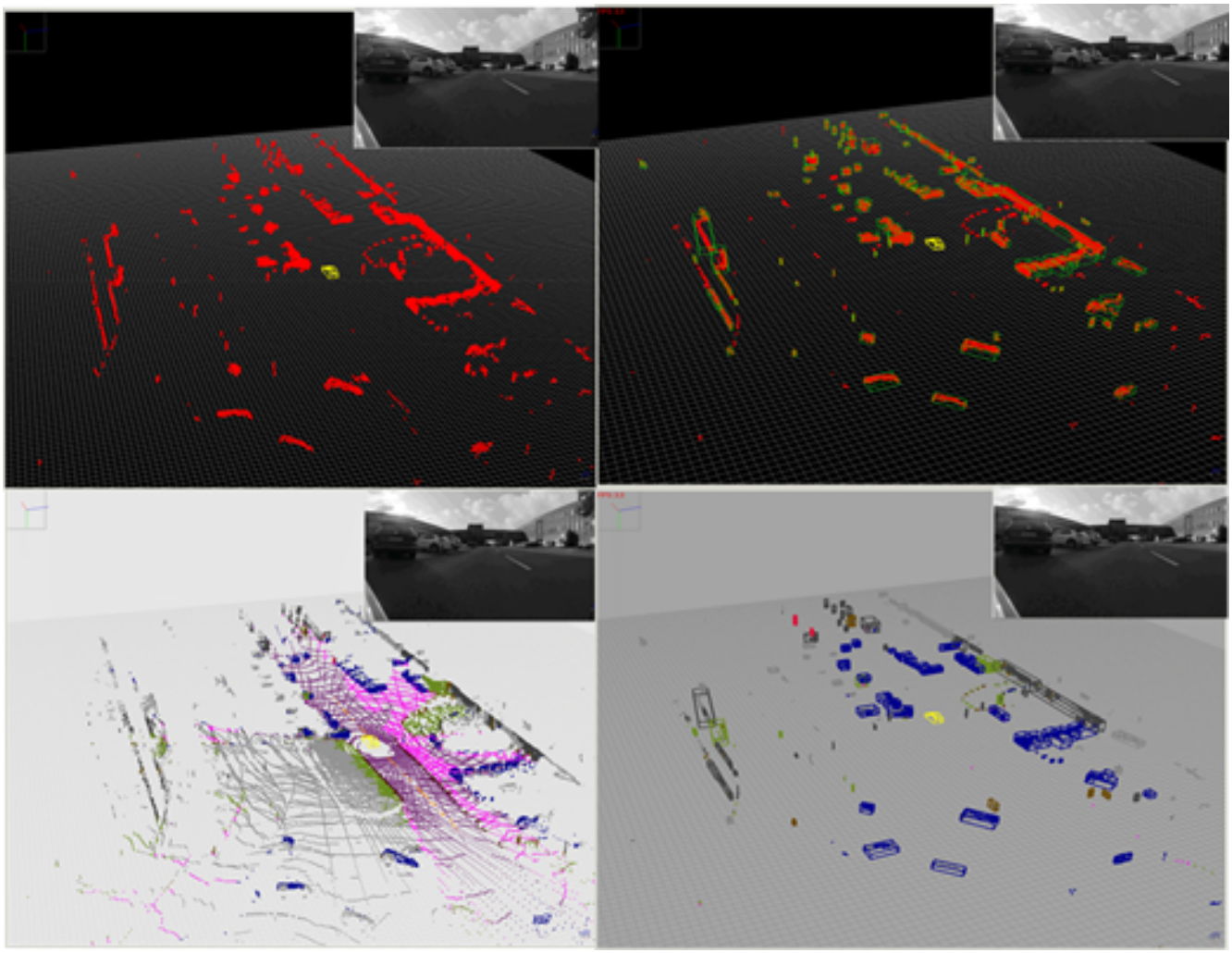


Fig. T4.1.5 Results from the paper “Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion”. Enhanced perception based on the Spatio-Temporal and Appearance Based Representation (STAR)

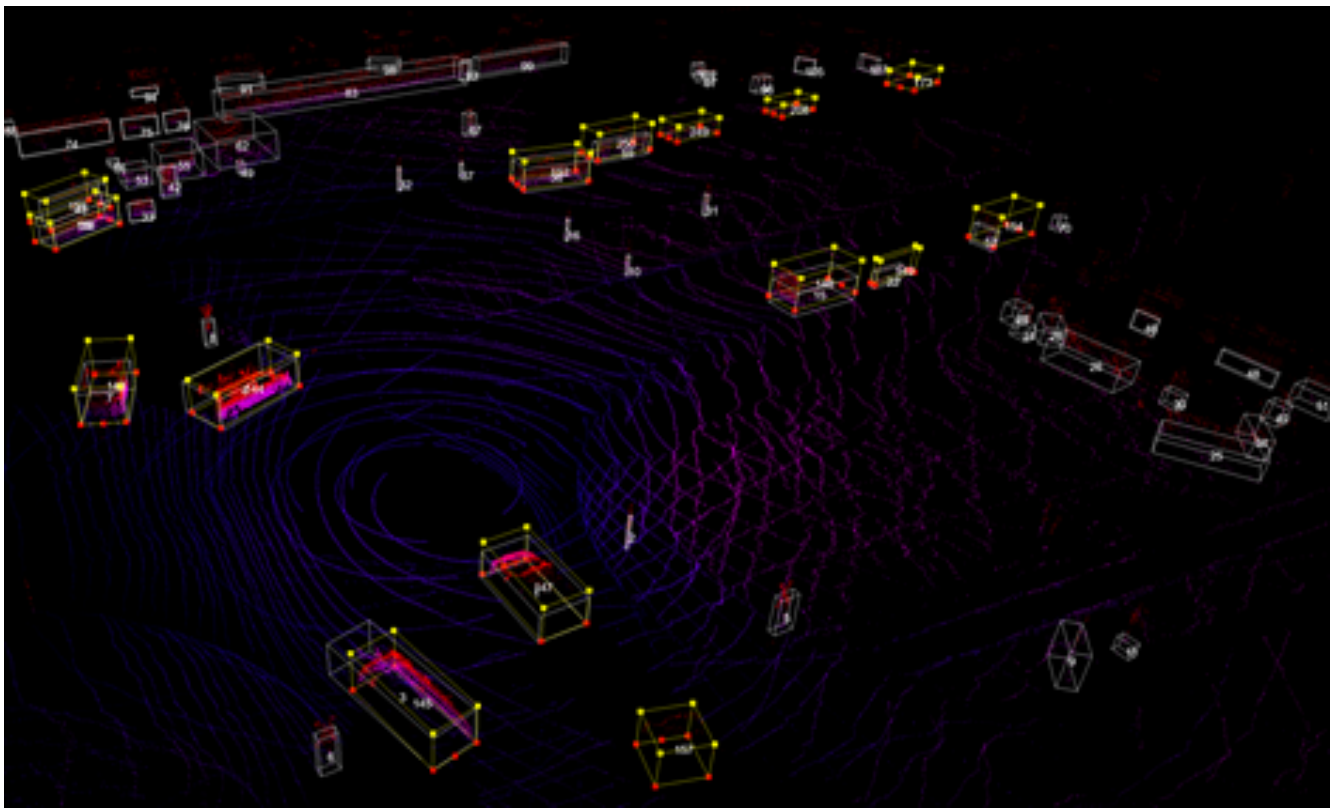


Fig. T4.1.6 Results from the paper “Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion”, Evaluation of detection, classification and localization by comparing 3D perception results with the 3D annotation of the acquired multimodal point cloud.

Reference

H. Florea, A. Petrovai, I. Giosan, F. Oniga, R. Varga, S. Nedevschi, “Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion”, submitted to IEEE Transactions on Intelligent Transportation Systems