Scientific - Technical Report

(2018 - 2020)

Competition:	**Complex Ground-Breaking Research Projects_PCCF 2016**
No. of the contract:	9/2018
Research domain:	PE6_7 - Artificial intelligence, intelligent systems, multi agent systems PE6_8 - Computer graphics, computer vision, multimedia, computer games
Title:	**Integrated Semantic Visual Perception and Control for Autonomous Systems**
Acronym:	SEPCA
Duration (months):	48
Total funded budget:	8.500.000,00
- 2018 settled	1.193.832,00
- 2019 settled	2.564.320,00
- 2020 settled	1.697.597,00
Project web page:	http://109.101.234.42/projects.php?ID_p=15
Host institution:	„Simion Stoilow" Institute of Mathematics of the Romanian Academy
Project manager	Cristian Sminchisescu
Coordinator (CO):	„Simion Stoilow" Institute of Mathematics of the Romanian Academy (IMAR)
Project partner 1 (P1):	Technical University of Cluj-Napoca (UTCN)

1. Summary of the context and overall objectives of the project (for the 2018-2020 period, include the conclusions of the action)

The objective of this project is to develop principled mathematical, computational and systems components in order to construct the next generation of autonomous vehicles capable of integrated visual perception (scene reconstruction and recognition) and action (planning and navigation) based on computer vision, machine learning, and optimal control techniques. A central contribution of this work is the development of fully trainable, large scale semantic architectures based on deep neural networks that enable the complete, end-to-end, training of the geometric, categorization and navigation parameters of the model in a single optimization process. By integrating and advancing components within computer vision, machine learning and optimal control, we will be able to develop perceptual robotics systems that can semantically map, navigate, and interact with an unknown environment. For demonstration, we will develop an autonomous system for the visual inspection of a forest using small UAVs (quadcopters), including classifying different types of trees, estimating their age and counting their numbers based on geometric and semantic information, as well as avoiding or following people. The demonstrator is interesting in its own right, but represents only a testbed for the methodology developed in the project, which is applicable broadly, to autonomous vehicles, humanoid robots, surveillance and security, or flexible inspection in general.

The 2018-2020 period

During the period under scrutiny we have addressed 3d geometrical modeling in simple scenarios and achieved several results in terms of deep learning of graph matching and deep structured geometric models with semantics. The results were published in top level computer vision journals and conferences as can be noticed in the section 4 devoted to results indicators. The topics that have been tackled range from integrated 3d sensing of multiple people in natural images, human synthesis and scene compositing and 3d reconstruction of human interactions, including 3d human self-contact learning, to real-time semantic segmentation based stereo reconstruction, semi-global optimization for classification-based monocular depth estimation or unified methods for improving long-range accuracy of stereo and monocular depth estimation algorithms. A particular application for monocular depth estimation techniques for environment perception from drones has been developed as well.

Another objective of the period concerned visual learning and recognition in simple scenarios. To this end, we have also developed several weakly-supervised semantic models with multiple components and partial responses as well as active and adversarial learning structures and methods for dynamic data for visual recognition and localization. The results highlighted here (and subsequently dealt with further in the report), address various topics such as semantic segmentation learning for autonomous UAVs using simulators as well as real data, a critical evaluation of aerial datasets for semantic segmentation, methods to narrow the semantic gap between real and synthetic data, weakly supervised semantic segmentation learning on UAV video sequences, fusion schemes for semantic and instance-level segmentation, multi-task networks for Panoptic segmentation in automated driving, methods for efficient instance and semantic segmentation for automated driving, real-time Panoptic segmentation with prototype masks for automated driving, video semantic segmentation leveraging dense optical flow or semantic cameras for 360-degree environment perception in automated urban parking and driving.

An important goal of the reporting period concerned the development of semantic optimal control methods for planning and navigation that integrate geometric and semantic information, in an adaptive, perception-and-action setting. In this respect, we are showing in the report the results we have obtained using reinforcement learning methods for semantic synthesis of pedestrian locomotion, deep reinforcement learning for active human pose estimation, embodied visual active learning for semantic segmentation or self-supervised active triangulation for 3D human pose reconstruction. The methodology and models are general and we highlight in the report how they can be used, for instance, in proof-of-concept experiments illustrating the way the developed technology can be connected to an active drone observer to reconstruct 3d poses.

Last but not least, the reporting period has been used to perform data collection for forest navigation, labelling and analysis. Using the equipment purchased in the project, namely DJI Matrice 210 V2 RTK drones, we have created an aerial imaging dataset. It comprises high-resolution images alongside video sequences from multiple flights over forest and open terrains, for which accurate positioning data is available. We generated textured 3D meshes for each flight area, along with a 3D point cloud and a digital elevation model. 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.

All these results have been made possible by using the dedicated equipment foreseen in the proposal to achieve the goals of the project. Tender procedures have been staged to buy two DGX-1-like systems with 8 Tesla V100 GPUs each. These computing systems have proven instrumental in training the deep network models presented in this report. Similarly, two DJI Matrice 210 RTK drones have been purchased and used to collect forest data that serves as training data for the deep models currently being developed.

We consider that all of the results of the project so far, including methodologies and models, that we highlighted here and subsequently explain at length in the report are well aligned with the objectives of the project as stated in the proposal roadmap and plan of action. The publication record (listed in the result indicators section of the report) which includes papers published, accepted or submitted at high impact factor journals as well as over 20 articles published at some of the most prestigious conferences in the field, shows undoubtedly the quality of the work described in this report.

2. **Work performed from the beginning of the project to the end of the period covered by the report and main results achieved so far**

A1. Deep 3D Reconstruction

Task 1.1: Deep Learning of Graph Matching under Global Constraints

The problem of graph matching under node and pairwise constraints is fundamental in areas as diverse as combinatorial optimization, machine learning or computer vision, where representing both the relations between nodes and their neighborhood structure is essential. In an article published at CVPR 2018, we present an end-to-end model that makes it possible to learn all parameters of the graph matching process, including the unary and pairwise node neighborhoods, represented as deep feature extraction hierarchies. The challenge is in the formulation of the different matrix computation layers of the model in a way that enables the consistent, efficient propagation of gradients in the complete pipeline from the loss function, through the combinatorial optimization layer solving the matching problem, and the feature extraction hierarchy. Our computer vision experiments and ablation studies on challenging datasets like PASCAL VOC keypoints, Sintel and CUB show that matching models refined end-to-end are superior to counterparts based on feature hierarchies trained for other problems. For qualitative results, fig. T1.1.1 below shows matching results obtained on the PASCAL VOC dataset. Qualitative results on the CUB dataset, including ground truth results are shown in fig. T1.1.2.

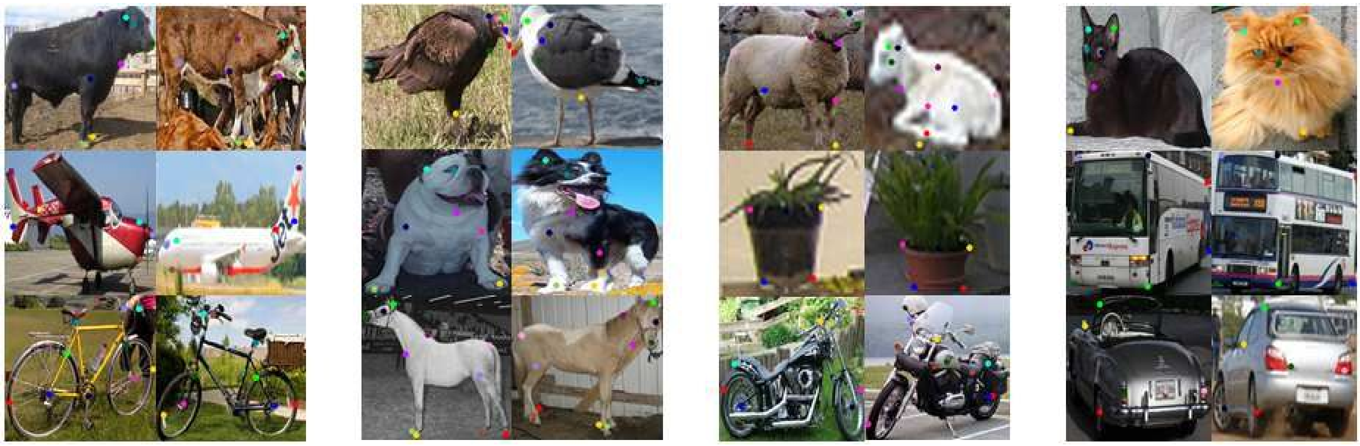


Fig. T1.1.1 Twelve qualitative examples of our best performing network on the PASCAL VOC test-set. For every pair of examples, the left shows the source image and the right the target. Colors identify the computed assignments between points. The method finds matches even under extreme appearance and pose changes.

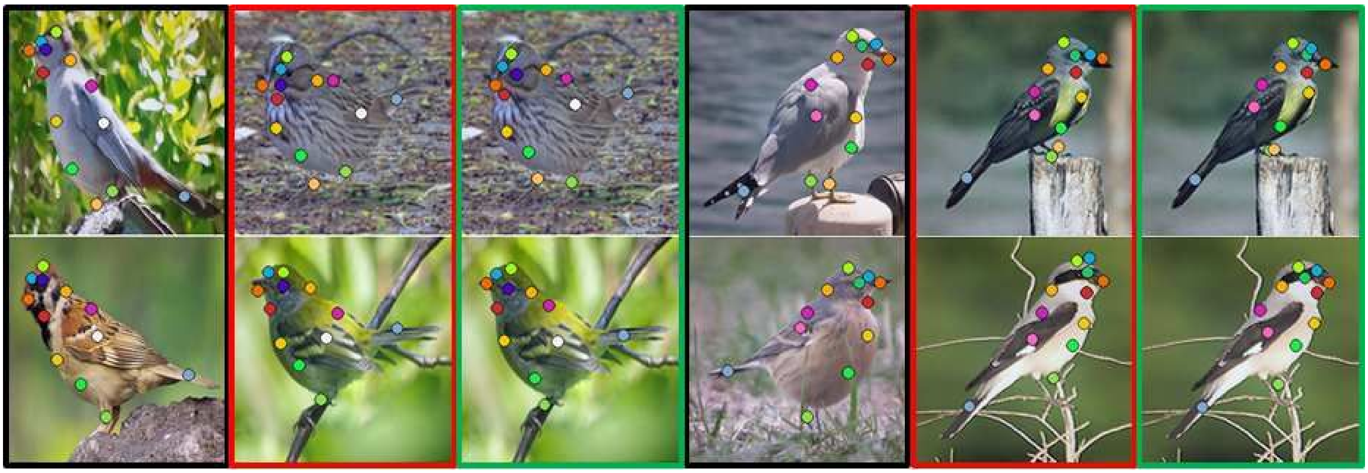


Fig. T1.1.2 Four qualitative examples of our best performing network on the CUB-200-2011 test-set. Images with a black contour represent the source, whereas images with a red contour represent targets. Color-coded correspondences are found by our method. The green framed images show ground-truth correspondences. The colors of the drawn circular markers uniquely identify 15 semantic keypoints.

Reference

A. Zanfir and C. Sminchisescu. "Deep Learning of Graph Matching", Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018

Task 1.2: Deep Structured Geometric Models with Semantics

Towards the goals we developed MubyNet – a feed-forward, multitask, bottom up system for the integrated localization, as well as 3d pose and shape estimation, of multiple people in monocular images. The challenge is the formal modeling of the problem that intrinsically requires discrete and continuous computation, e.g. grouping people vs. predicting 3d pose. The model identifies human body structures (joints and limbs) in images, groups them based on 2d and 3d information fused using learned scoring functions, and optimally aggregates such responses into partial or complete 3d human skeleton hypotheses under kinematic tree constraints, but without knowing in advance the number of people in the scene and their visibility relations. We design a multi-task deep neural network with differentiable stages where the person grouping problem is formulated as an integer program based on learned body part scores parameterized by both 2d and 3d information. This avoids suboptimality resulting from separate 2d and 3d reasoning, with grouping performed based on the combined representation. The final stage of 3d pose and shape prediction is based on a learned attention process where information from different human body parts is optimally integrated. State-of-the-art results are obtained in large scale datasets like Human3.6M and Panoptic, and qualitatively by reconstructing the 3d shape and pose of multiple people, under occlusion, in difficult monocular images. Qualitative results are shown below in fig. T1.2.1.

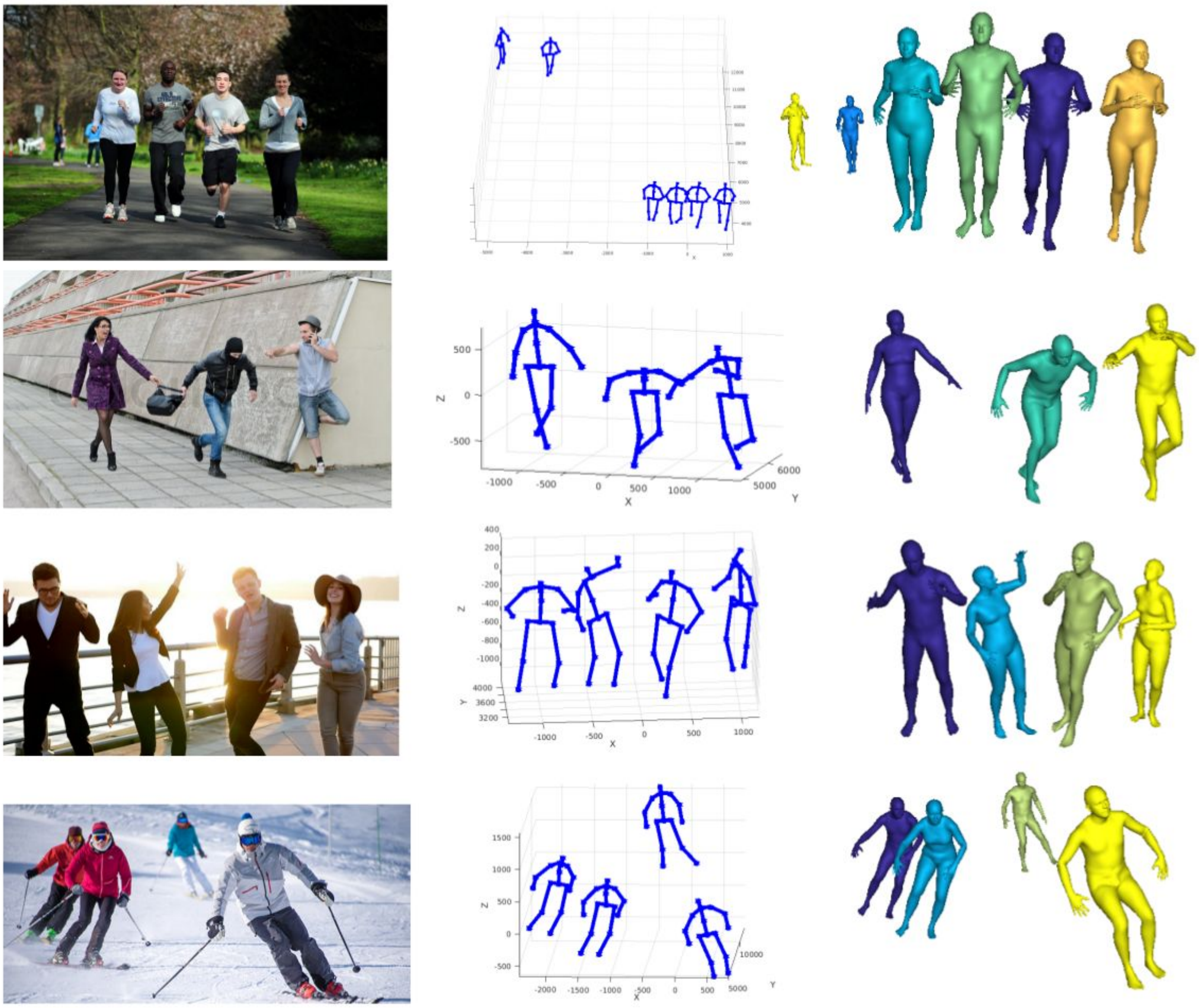


Fig T1.2.1 Human pose and shape reconstruction of multiple people produced by MubyNet illustrate good 3d estimates for distant people, complex poses or occlusion. For global translations, we optimize the Euclidean loss between the 2d joint detections and the projections predicted by our 3d models.

Reference

A. Zanfir, E. Marinoiu (Oneata), M. Zanfir, A. Popa, C. Sminschisescu. "Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images", Proceedings of the Thirty-second Conference on Neural Information Processing Systems, NIPS 2018

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.

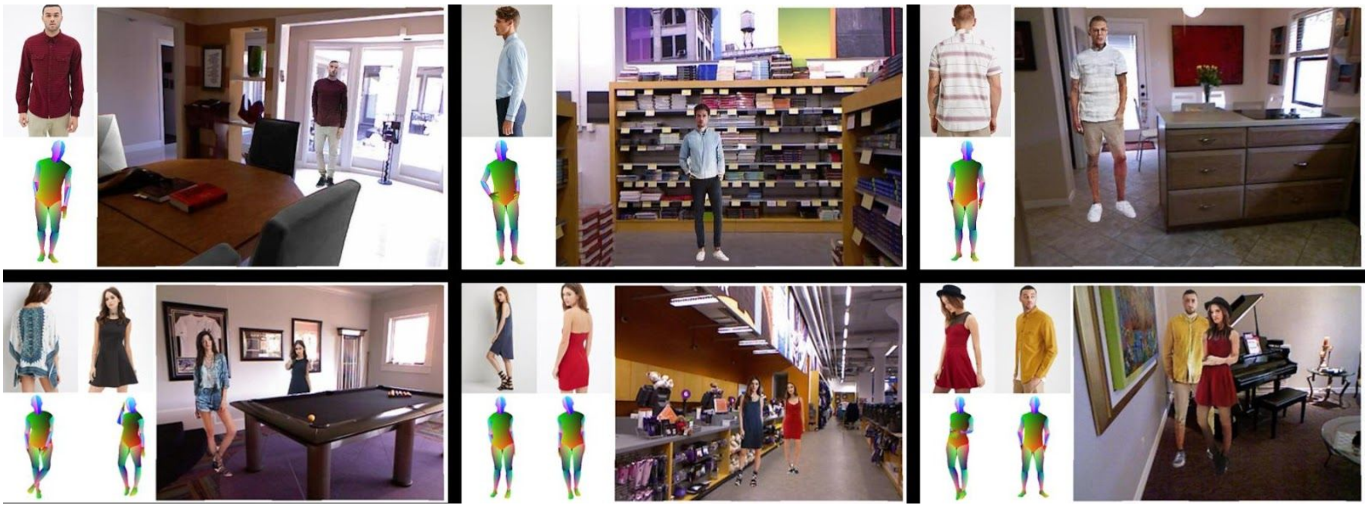


Fig T1.2.2 Sample images generated by our framework. For each example, we show the source image, the target 3d body mesh, and a scene with a geometrically plausible placement of the synthesized person. Please note that our framework allows for a positioning behind various objects, and the insertion of 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.

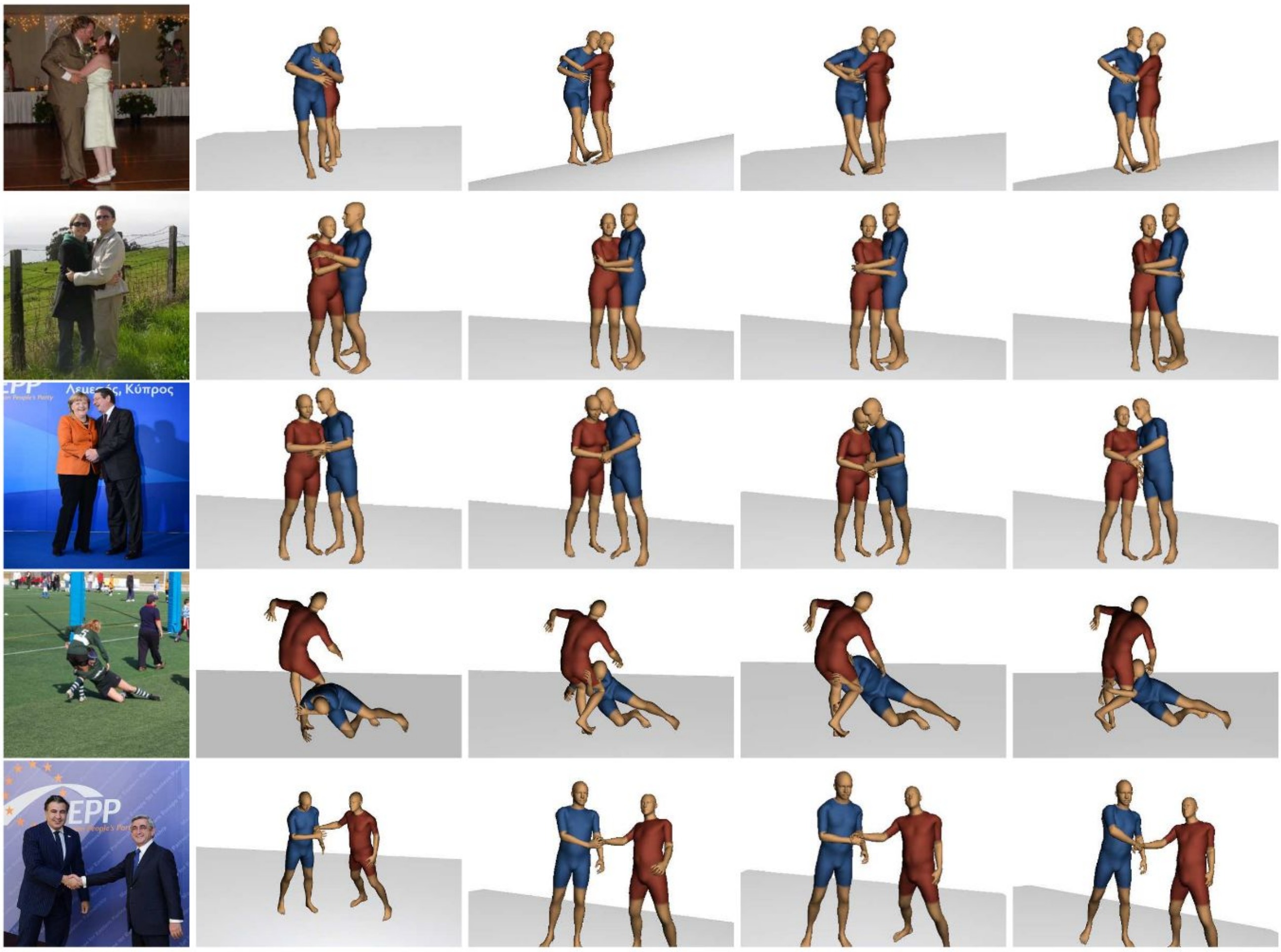


Fig. T1.2.3 3D human pose and shape reconstructions using contact constraints of different granularity. The first column shows the RGB images followed by their reconstructions without contact information (column 2), using contacts based on 37 and 75 regions, respectively (columns 3 & 4), and using facet-based 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.

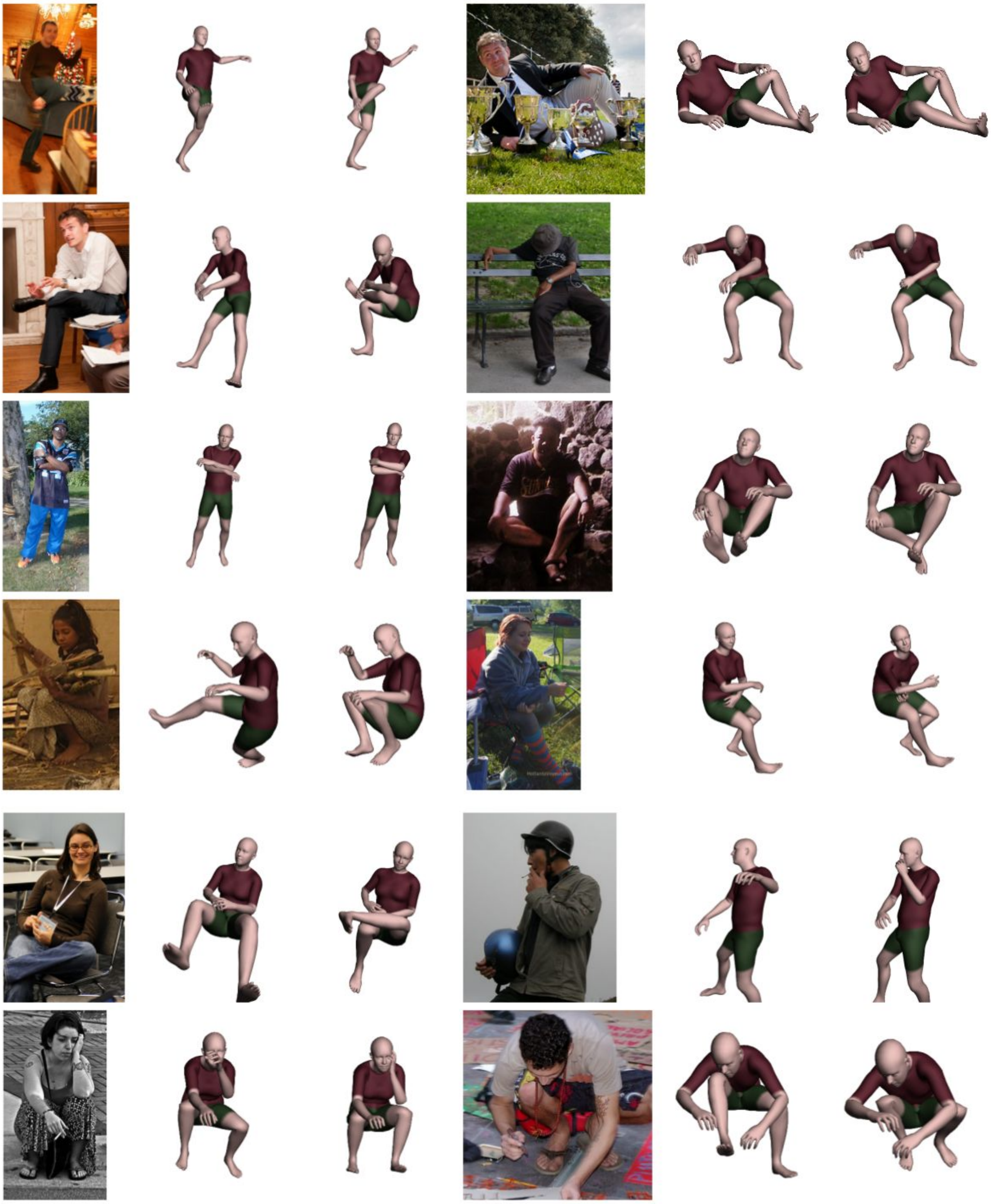


Fig. T1.2.4 3D pose and shape reconstructions using our annotated self-contact data. Left: Original image. Center: Reconstruction without considering the self-contact and the associated loss. Right: Reconstruction that uses the self-contact annotations and the corresponding loss.

Reference

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

In the larger context of scene understanding and aligned with the objectives of the project we also would like to show some qualitative results for semantic segmentation of forest environments obtained with a drone purchased in the project, as part of the effort to build datasets with forest images. Fig. T1.2.5 depicts semantic segmentation of deforestation areas from aerial images as well as instance level segmentation of trees within the forest.

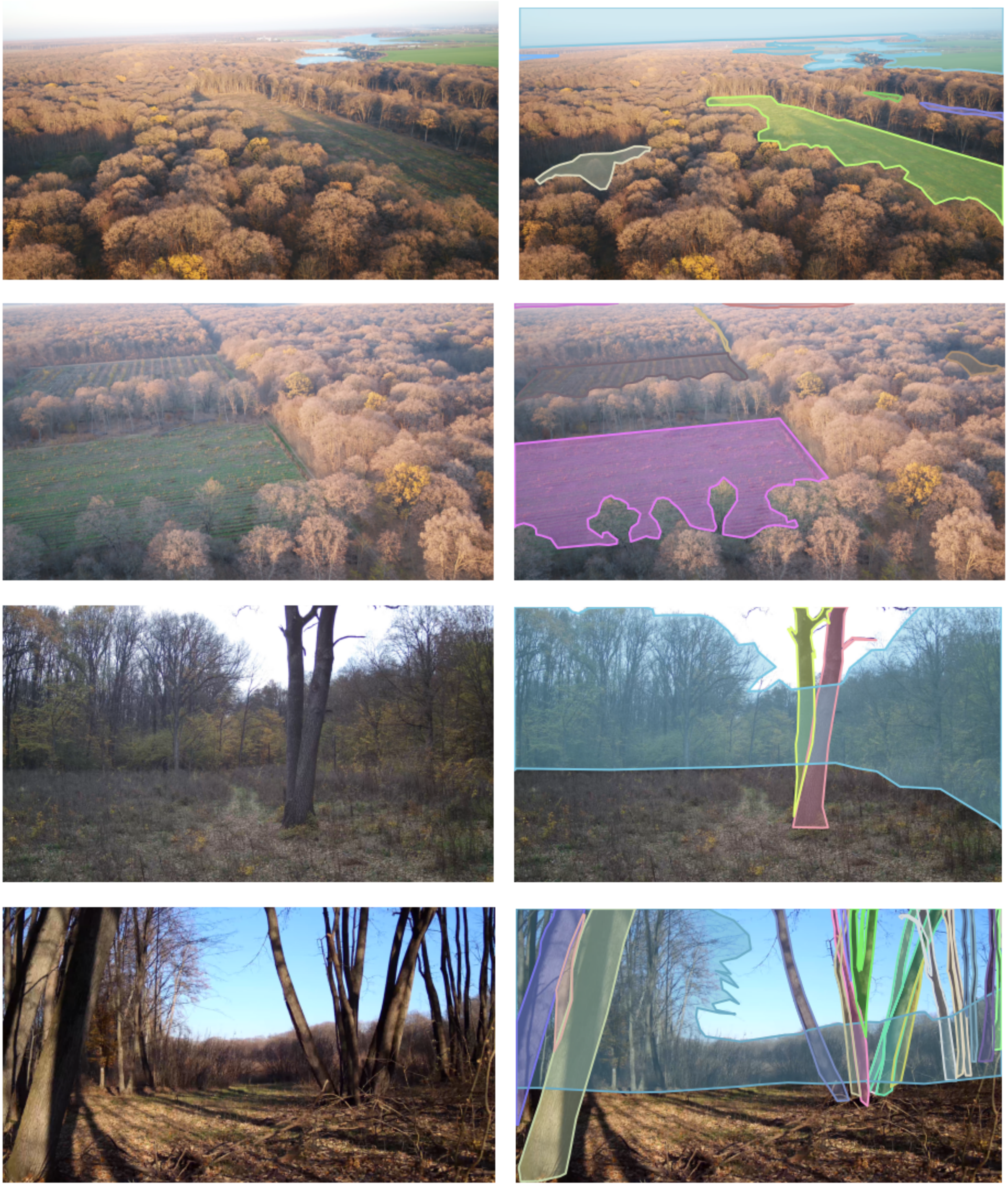


Fig. T1.2.5 Aerial images with semantic segmentation of the non-forest areas (top two rows) and forest level images with instance segmentations of the trees and semantic segmentation of the forest (bottom two rows. Images on the left column represent the original record taken by the drone, while the left column presents the semantic segmentation result.

Task 1.2 deals with producing a 3D representation of the environment, in which each 3D point is also associated with semantic class. A contribution with respect to this goal "Real-Time Semantic Segmentation-Based Stereo Reconstruction" was done by generating the depth map (first step required for the 3D representation) by enhancing the stereo reconstruction process with semantic information. To this end, initially a semantic map of the scene is generated by using a convolutional neural network. Then, each sub-task of the stereo reconstruction algorithm is tailored to incorporate scene information obtained from the semantic map and thus to enhance the results. New learning algorithms (based on genetic algorithms and convolutional neural networks) are introduced for these steps. Results show that the new method produced the best real-time stereo reconstruction results on the Kitti stereo benchmark. Although using stereo reconstruction on aerial images is both cumbersome and susceptible to errors (the system can easily decalibrate), this approach is really important because it demonstrates the benefits of using high-level scene information (provided through the semantic map) for low-level vision tasks required for depth perception.

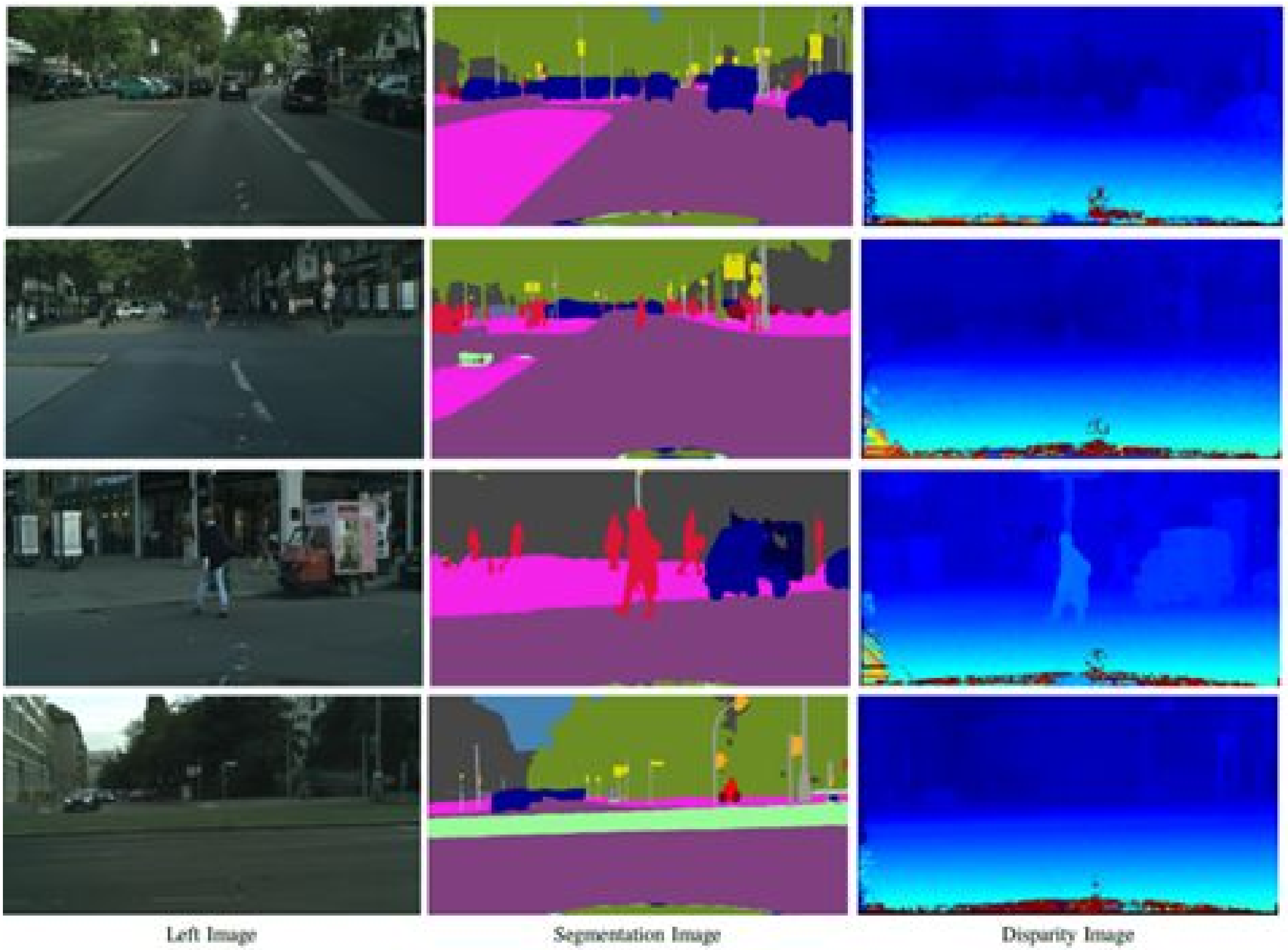


Fig. T1.2.6 Results from the paper "Real-Time Semantic Segmentation-Based Stereo Reconstruction". The disparity image (from Cityscapes dataset) is obtained by enhancing the stereo pipeline with information from the semantic segmentation. The proper shape of the objects is clearly captured in all scenarios

Reference

V. Miclea, S. Nedevschi, "Real-Time Semantic Segmentation-Based Stereo Reconstruction", IEEE Transactions on Intelligent Transportation Systems (Early Access), pp. 1-11, 2019

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.

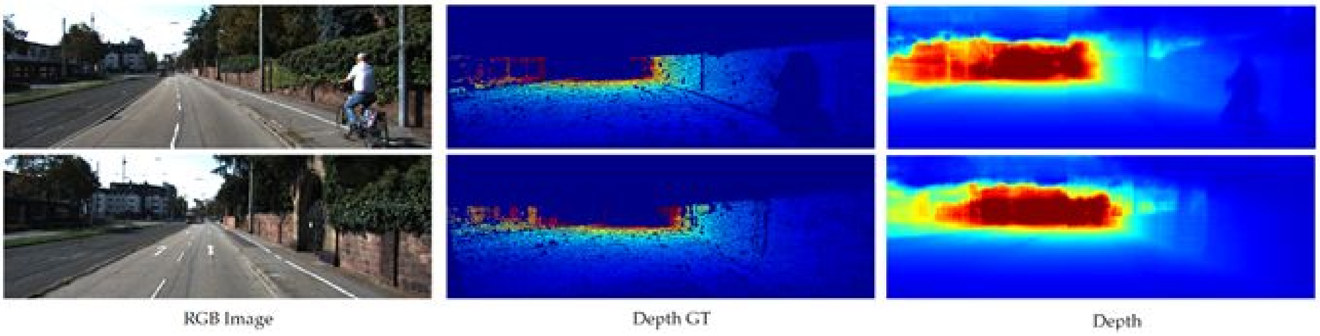


Fig. T1.2.7 Results from the paper "Semi-Global Optimization for Classification-Based Monocular Depth Estimation". The depth map is accurately generated, by using the novel stereo-based global optimization. The results also present a higher confidence, due to the geometrical constraints introduced by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

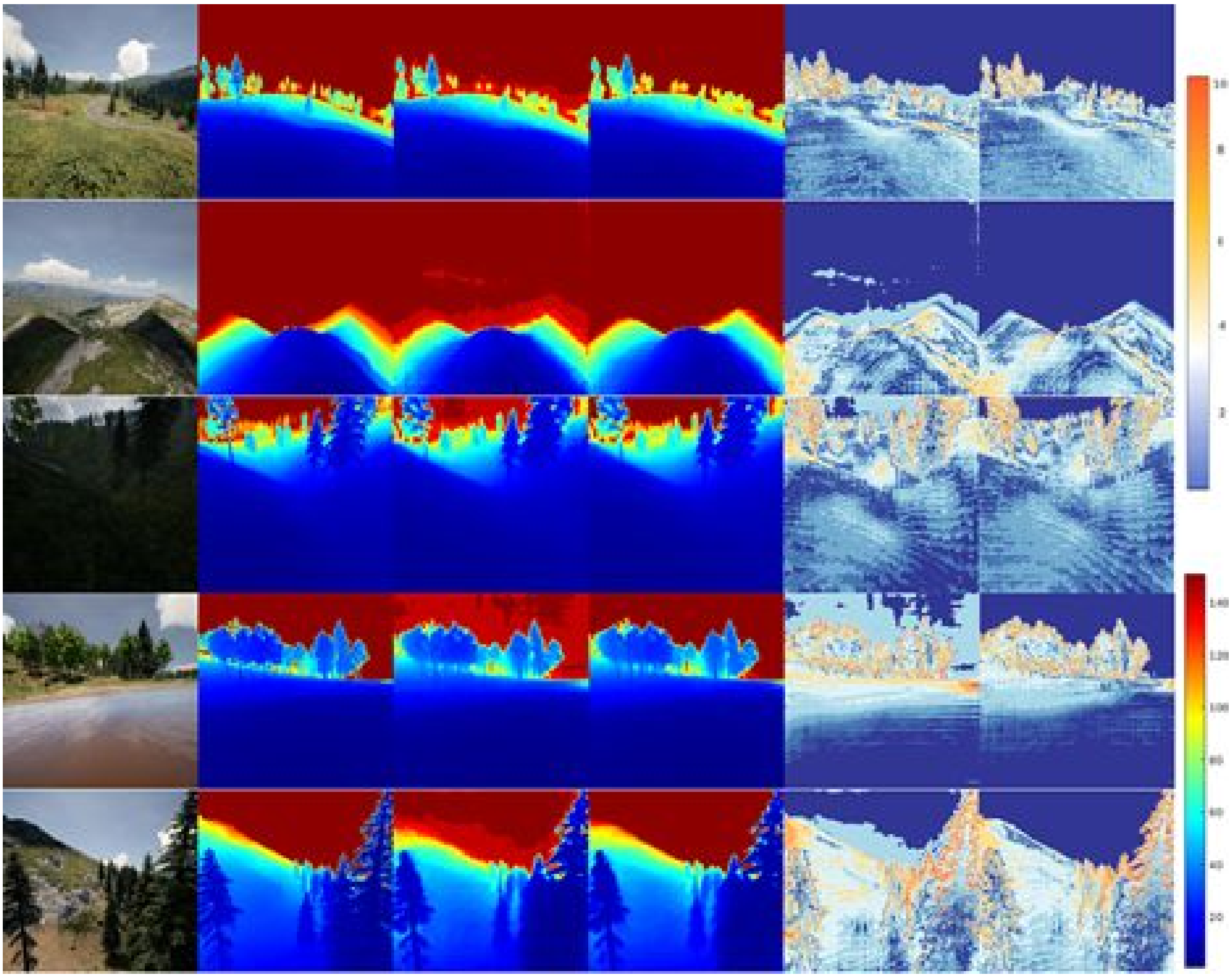


Fig. T1.2.8 Results from the article “A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithms”. MidAir images (synthetic) are used as input for the CNN, which produces highly accurate results (error images are presented on the last two columns)

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).

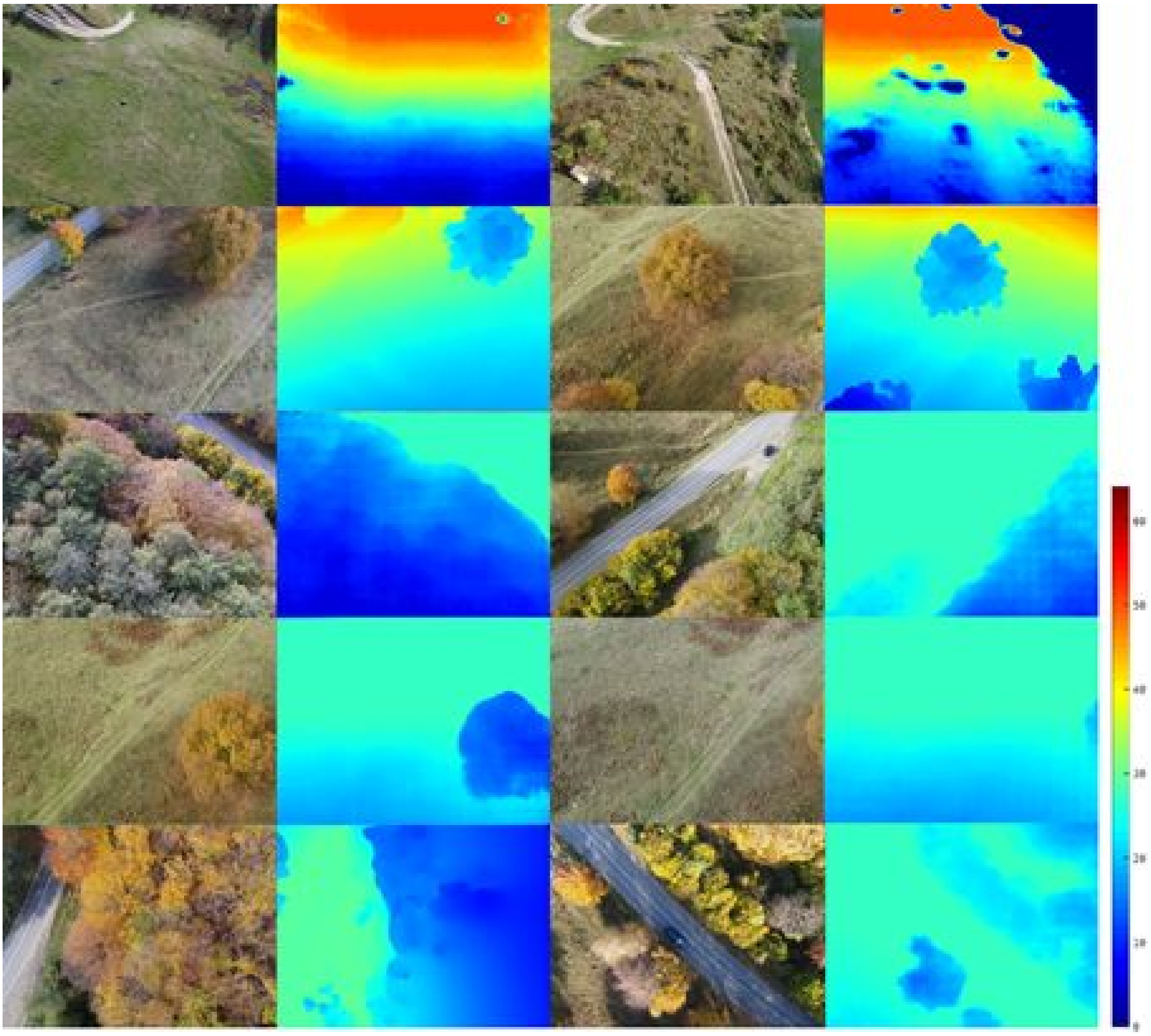


Fig. T1.2.9 Results from the article "Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception". The method produces very accurate depth results on real-life images (Nadir and at various other angles), captured from a real drone, in various scenarios (fields, forests)

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 order to solve the need of aerial annotated dataset for the multiple learning tasks we focused our attention on using synthetic dataset, transfer learning, reducing the semantic gap between synthetic and real data, and weakly-supervised semantic segmentation of video sequences.

In “Semantic Segmentation Learning for Autonomous UAVs using Simulators and Real Data” we made a thorough survey of five simulators (Gazebo, Udacity, Sim4CV, AirSim, and CARLA) and five synthetic datasets (SYNTHIA, Sintel, GTA V: Playing for Data, GTA V: Driving in the Matrix, and Virtual KITTI), exploring solutions for semantic segmentation on images taken from drones. We explored the problem of knowledge transfer by evaluating a deep learning model trained on both synthetic and real data (TUGRAZ drone dataset). We conclude that fine-tuning a large synthetic dataset with a smaller real one gives the best results.

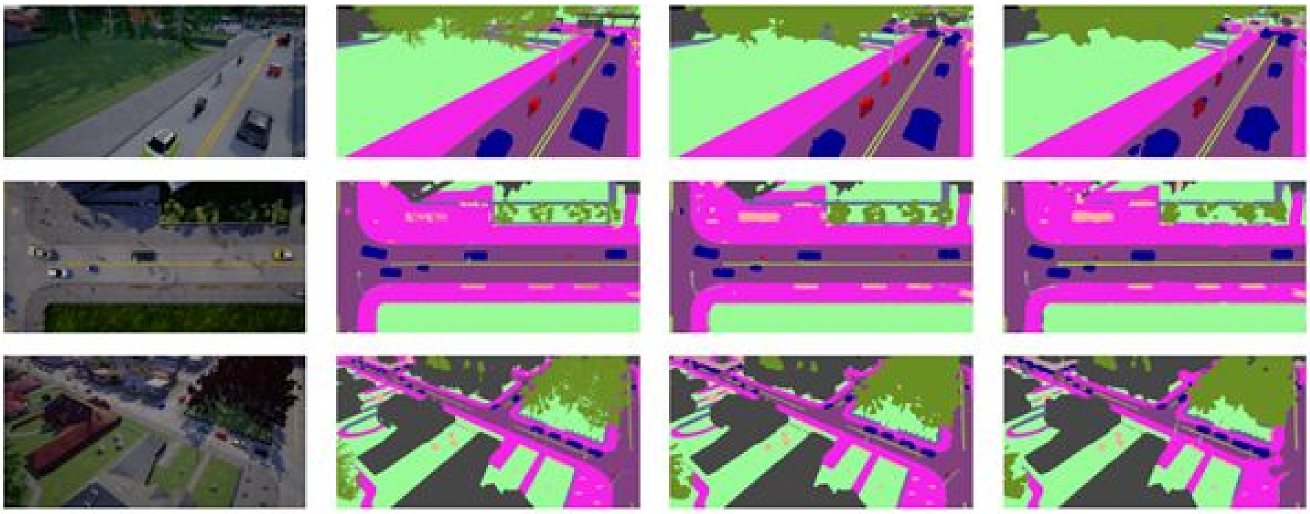


Fig. T2.1.1 Results from the paper **“**Semantic Segmentation Learning for Autonomous UAVs using Simulators and Real Data**”.** The evaluation of the semantic segmentation on the synthetic dataset. From left to right: RGB image, ground truth for semantic annotation, inferred image when network is trained on CARLA, inferred image when the network is trained on the merged dataset.

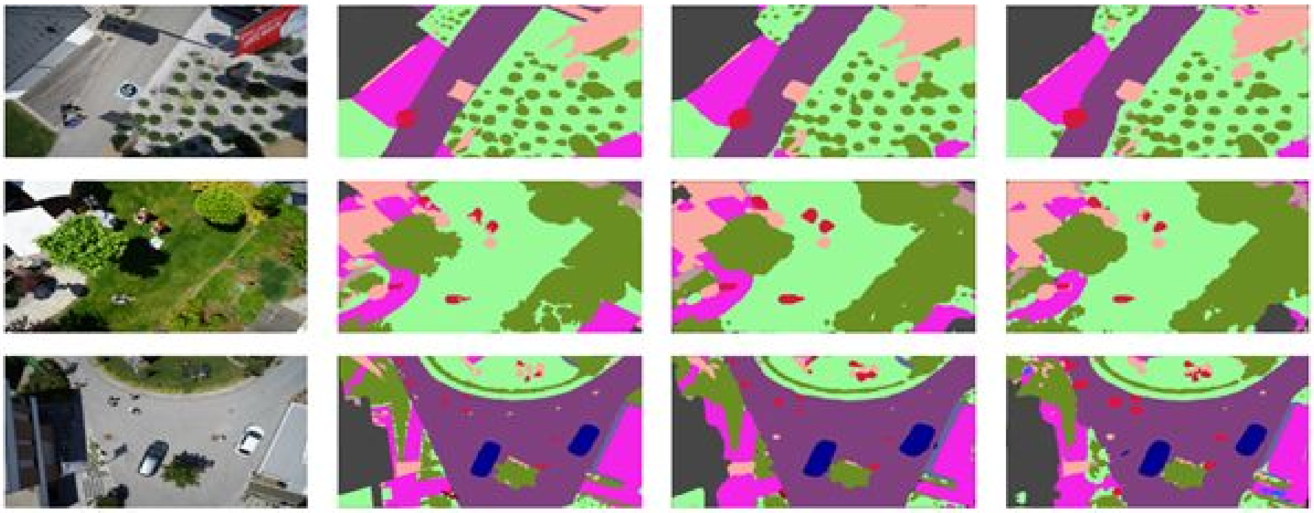


Fig. T2.1.2 Results from the paper **“**Semantic Segmentation Learning for Autonomous UAVs using Simulators and Real Data**”.** The evaluation of the semantic segmentation on the real drone dataset. From left to right: RGB image, ground truth for semantic annotation, inferred image when network is trained on the real dataset, inferred image when the network is trained on the merged dataset.

Reference

B. C. Z. Blaga, S. Nedevschi, „Semantic Segmentation Learning for Autonomous UAVs using Simulators and Real Data”, Proceeding of IEEE Intelligent Computer Communication and Processing (ICCP), 2019

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.

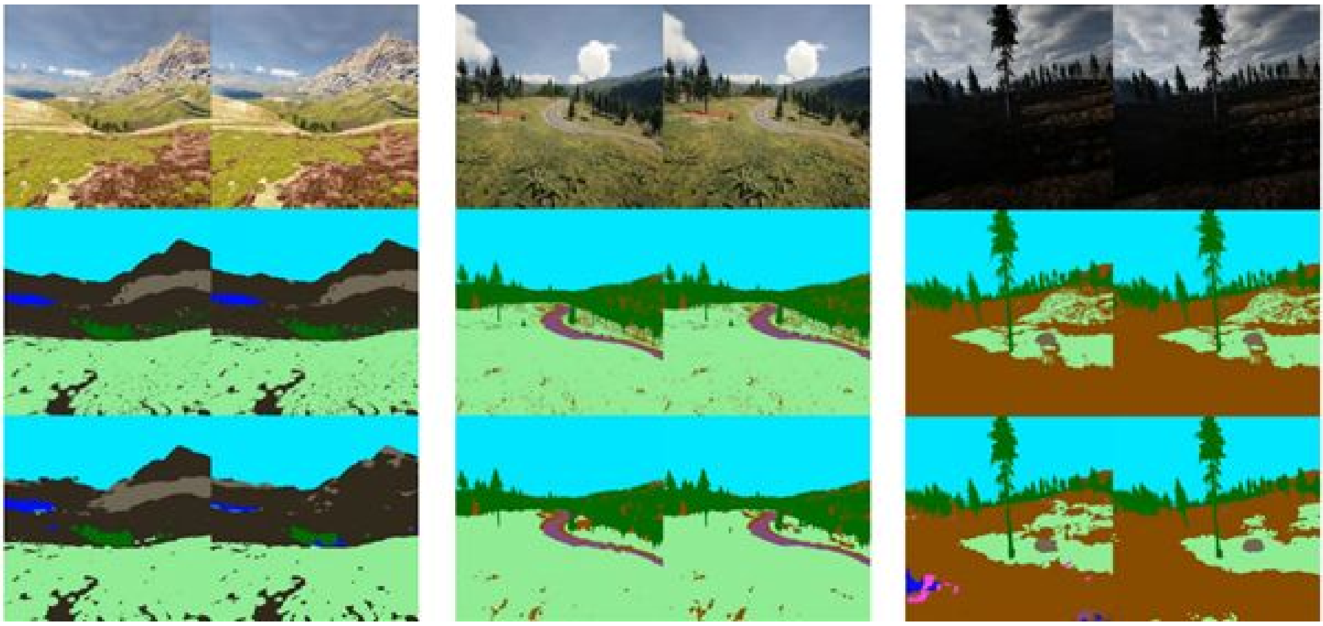


Fig. T2.1.3 Results from the paper ”A Critical Evaluation of Aerial Data Datasets for Semantic Segmentation”. Prediction results on the Mid-Air dataset, in 3 scenarios: mountain area, road in spring, and sunset in autumn. From top to bottom, the color image, the ground truth segmentation, and the semantic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

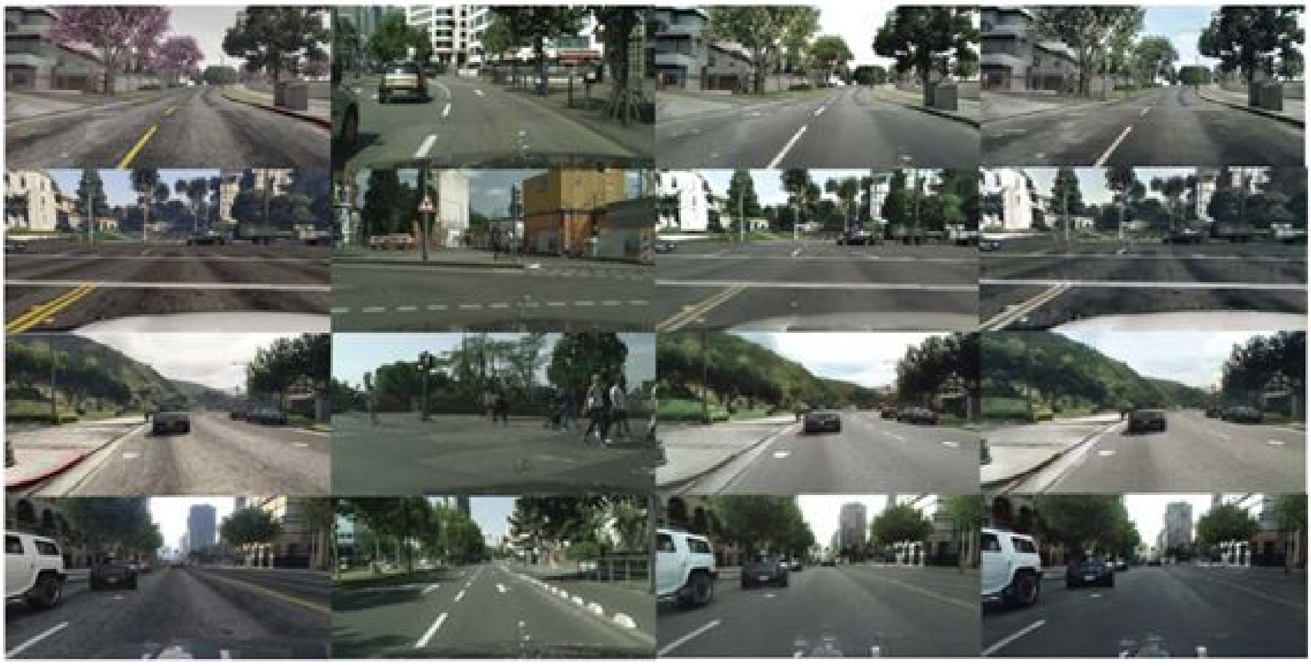


Fig. T2.1.4 Results from the paper “Narrowing the semantic gap between real and synthetic data”. Sample images from the translation of GTA5 → Cityscapes. From left to right: the first image represents the original synthetic image, the second image represent the style which was applied, the third column 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.

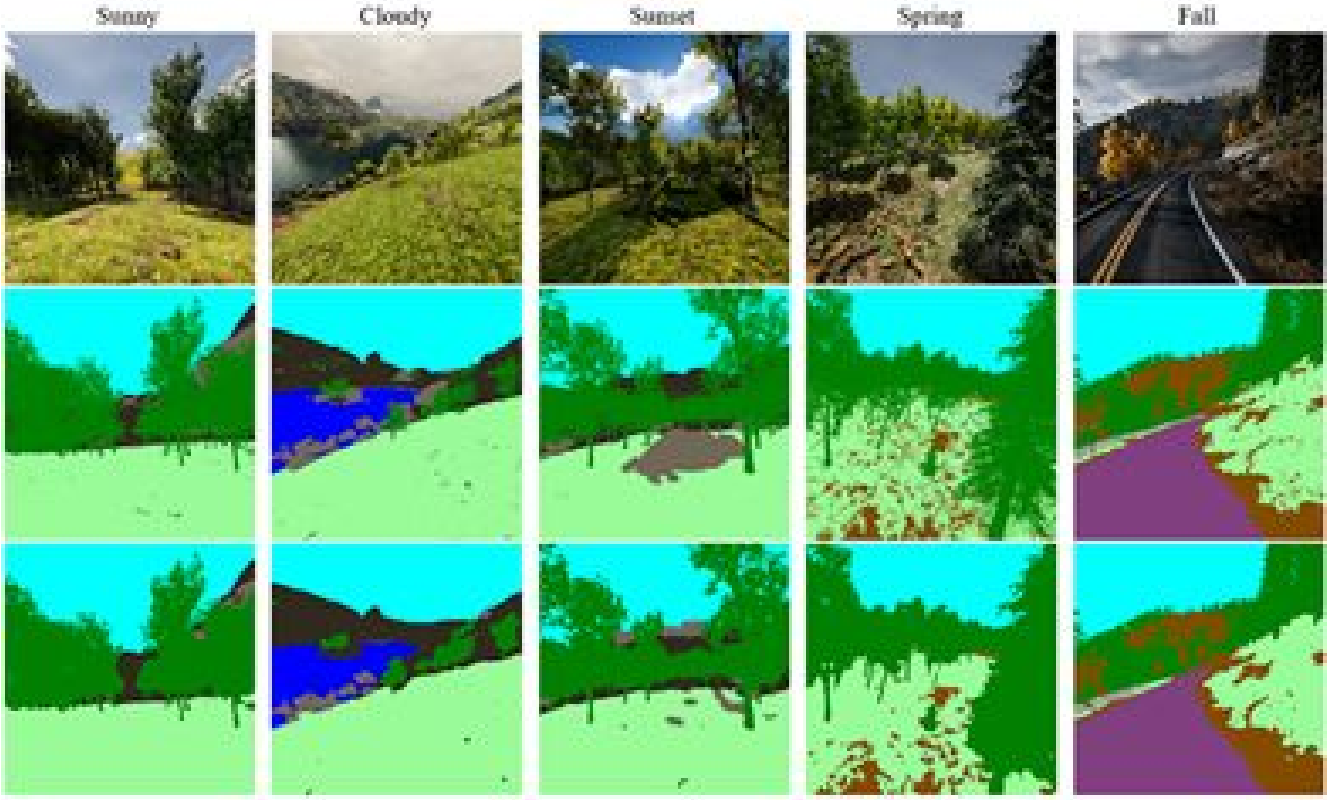


Fig. T2.1.5 Results of the paper “Weakly Supervised Semantic Segmentation Learning On UAV Video Sequences”. Results of the framework on the test set, presented for the 5 different scenarios from Mid-Air.

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.

This task focuses on the design and of computational procedures that are amenable to the large-scale training of dynamic data. During the course of the project, we studied, designed and implemented novel convolutional architectures for panoptic image segmentation. Panoptic segmentation provides pixel-level classification and instance identifiers for dynamic objects in the scene.

In our first solution ”Fusion Scheme for Semantic and Instance-Level Segmentation”, we propose a fusion scheme for instance and semantic segmentation based on heuristics to solve conflicts, which could be applied as a fast post-processing step on top of any semantic and instance network. We base our fusion module on the observation that semantic segmentation performs well in background segmentation, but struggles with foreground classes, where pixels of objects having different class but same category (for example truck, bus) are often misclassified (Fig. 1, column 3). In the case of instance segmentation, pixels of objects are correctly classified but the mask is more coarse than in the case of segmentation. We propose a post-processing step that provides a correction mechanism by propagating instance class and label on the semantic path at category level. We observe significant accuracy increases especially in the case of large objects.

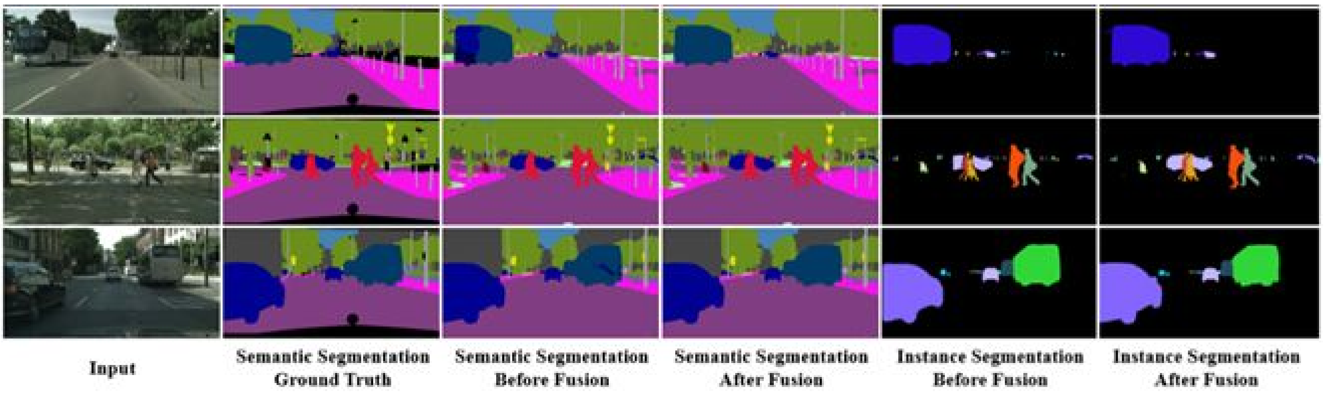

Input
Semantic Segmentation Ground Truth
Semantic Segmentation Before Fusion
Semantic Segmentation After Fusion
Instance Segmentation Before Fusion
Instance Segmentation After Fusion

Fig. T2.2.1 Results from the paper ”Fusion Scheme for Semantic and Instance-Level Segmentation”. The proposed post-processing fusion scheme yields a more accurate semantic segmentation, especially in the case of large objects.

Reference

A. D. Costea, A. Petrovai, S. Nedevschi, ”Fusion Scheme for Semantic and Instance-Level Segmentation”, Proceedings of 2018 IEEE Intelligent Transportation Systems Conference (ITSC), Maui, Hawaii, USA, November 4-7, 2018.

Next, we introduce a panoptic head which is end-to-end trainable with the multi-task semantic and instance segmentation network in „Multi-Task Network for Panoptic Segmentation in Automated Driving”. The panoptic head performs semantic and instance level recognition by pixel-level classification. Panoptic logits corresponding to background classes are built from the semantic segmentation logits, which are refined using instance masks from the instance segmentation head. Object mask logits from the instance segmentation head are as well improved by employing a sampling procedure at category level guided by the semantic foreground segments. Extensive experiments on the large-scale Cityscapes dataset shows that the proposed refinements of the semantic and instance masks and learning the panoptic output in an end-to-end manner brings significant accuracy gains to all tasks.

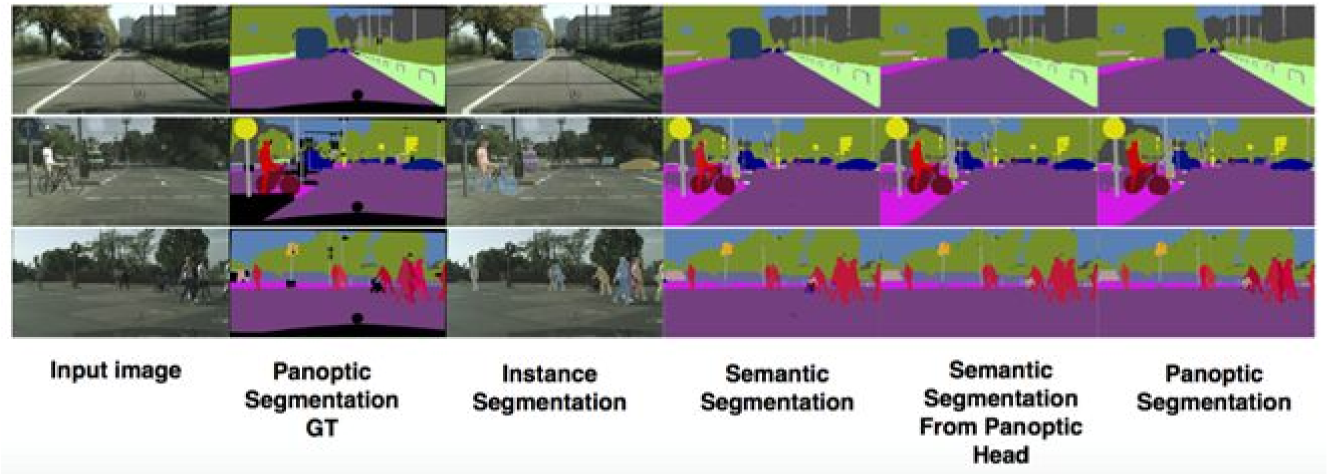


Fig. T2.2.2 Results from the paper „Multi-Task Network for Panoptic Segmentation in Automated Driving”. End-to-end learning of, instance and semantic segmentation improves both semantic and instance segmentation results.

Reference

A. Petrovai, S. Nedevschi, „Multi-Task Network for Panoptic Segmentation in Automated Driving", Proceeding of 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 26-30 October, 2019

Real-time performance is crucial for many robotic applications that perform online environment perception. Although these two-stage semantic and instance segmentation methods are accurate, they are not suitable for real-time processing. In the paper "Efficient instance and semantic segmentation for automated driving"we study how to speed up two-stage semantic and instance networks and propose a fast and efficient two-stage network that can reach better accuracy than the slower baseline. Our proposed network features a backbone with factorized convolutions and dilated convolutions for increased accuracy.

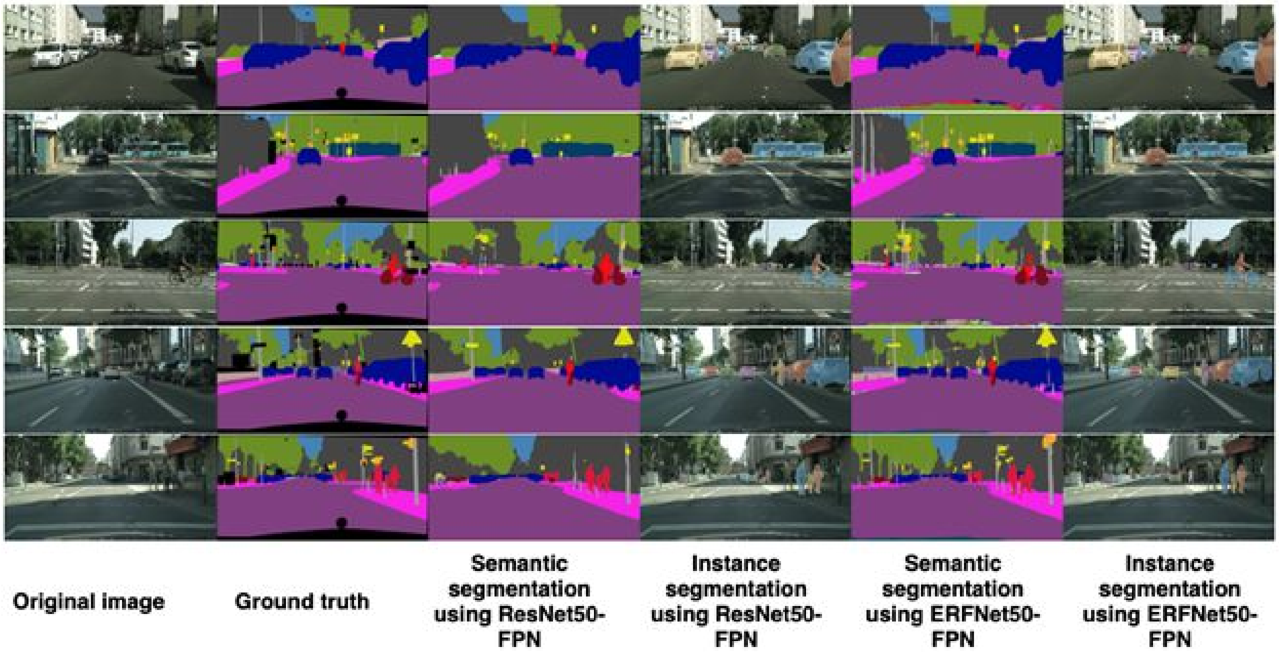

Original image
Ground truth
Semantic segmentation using ResNet50-FPN
Instance segmentation using ResNet50-FPN
Semantic segmentation using ERFNet50-FPN
Instance segmentation using ERFNet50-FPN

Fig. T2.2.3 Results from the paper "Efficient instance and semantic segmentation for automated driving". The ResNet50-FPN is the baseline network and the ERFNet50-FPN is the proposed network. Compared to the baseline, the proposed solution increased the segmentation mIoU with 4.5%.

Reference

A. Petrovai, S. Nedevschi, "Efficient instance and semantic segmentation for automated driving", Proceeding of 2019 IEEE Intelligent Vehicles Symposium (IV 2019), Paris; France; 9 - 12 June, 2019.

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.

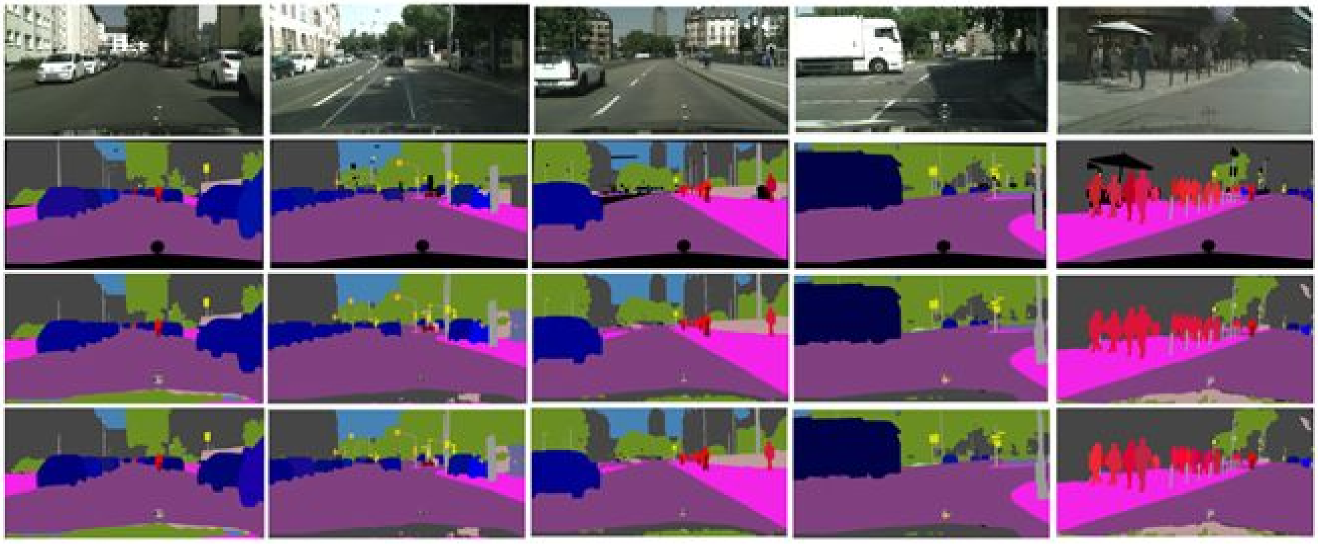


Fig. T2.2.4 Results from the paper "Real-Time Panoptic Segmentation with Prototype Masks for Automated Driving". From top to bottom: image, panoptic ground truth, semantic segmentation, panoptic segmentation. In the panoptic segmentation the color encodes the class and the instance identifier. Our network 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.

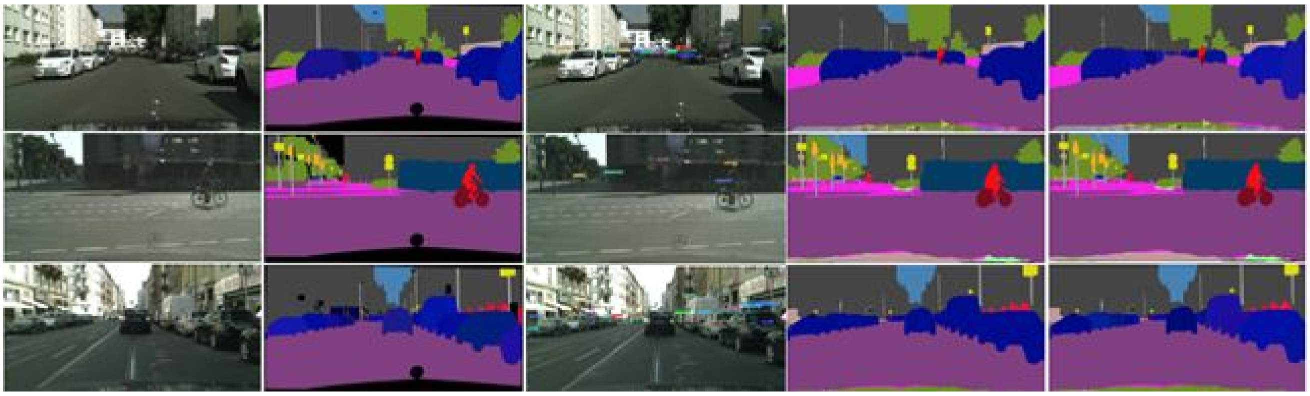


Fig. T2.2.5 Results from the paper "SAPSNet: A Soft Attention Panoptic Segmentation". From left to right: image, panoptic segmentation ground truth, object detection, semantic segmentation and panoptic segmentation. Our network can accurately segment object of various sizes and can handle difficult scenarios 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.

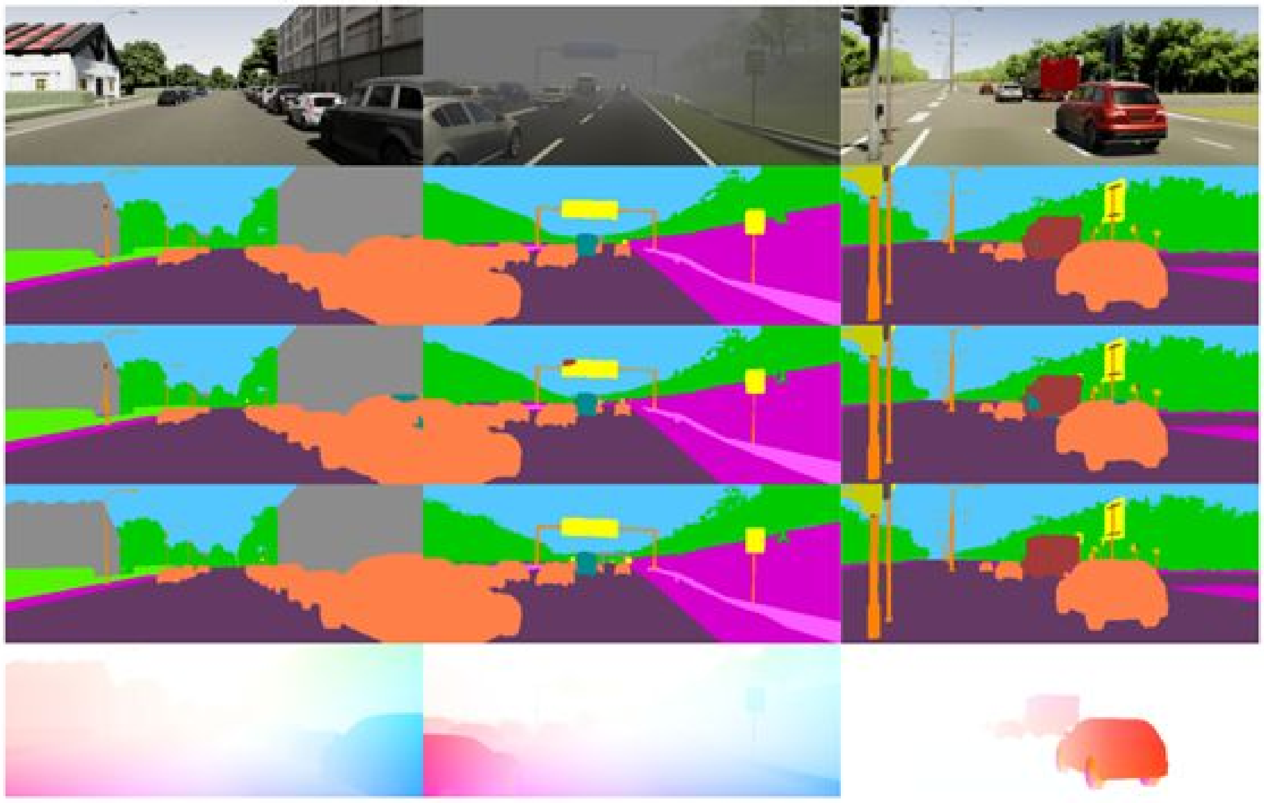


Fig. T2.2.6 Results from the paper "Video Semantic Segmentation leveraging Dense Optical Flow". Examples of segmentation on Virtual KITTI test subset. From top to bottom are the current frame, refined segmentation, static segmentation, ground truth and backward optical flow.

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.

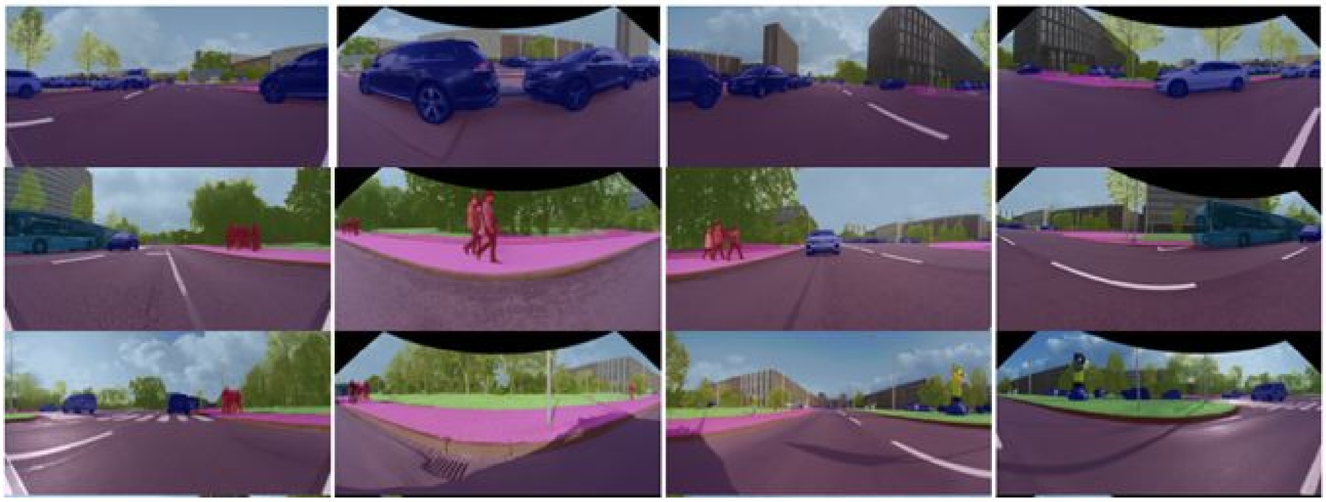


Fig. T2.2.7 Results from the paper “**Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving**”. Semantic segmentation of unwarped fisheye images. We process four images from the fisheye 160◦ horizontal field-of-view cameras which provide 360◦ coverage around the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

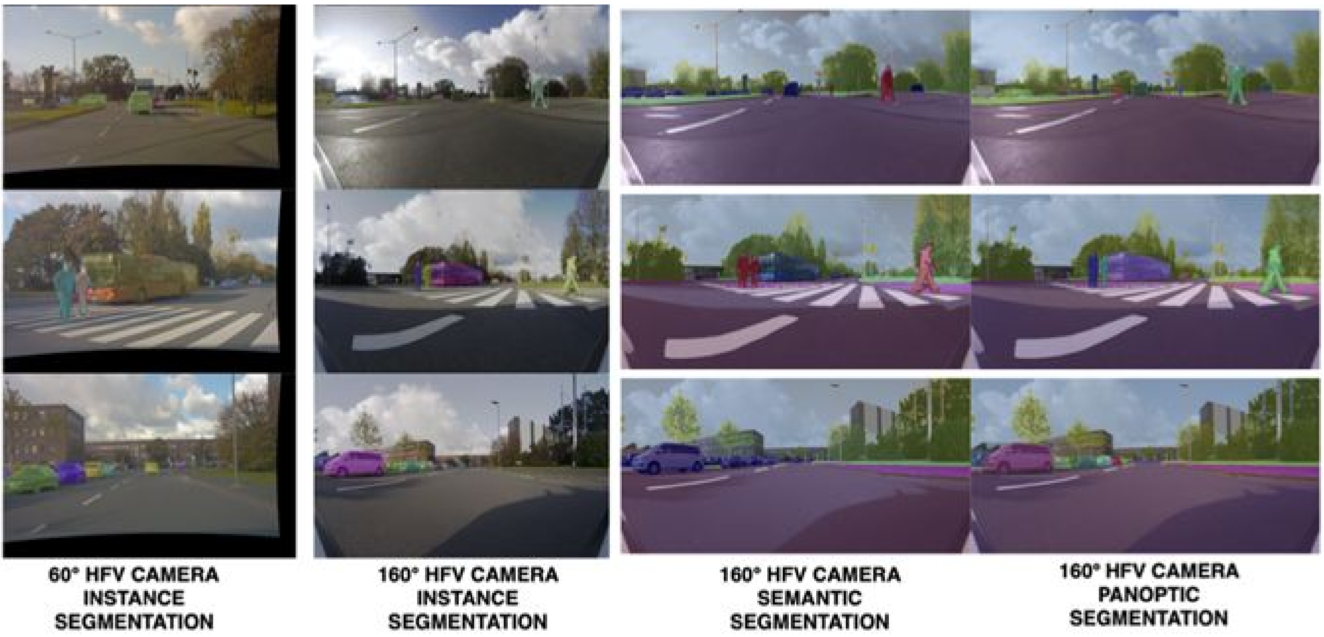

60° HFV CAMERA INSTANCE SEGMENTATION
160° HFV CAMERA INSTANCE SEGMENTATION
160° HFV CAMERA SEMANTIC SEGMENTATION
160° HFV CAMERA PANOPTIC SEGMENTATION

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.

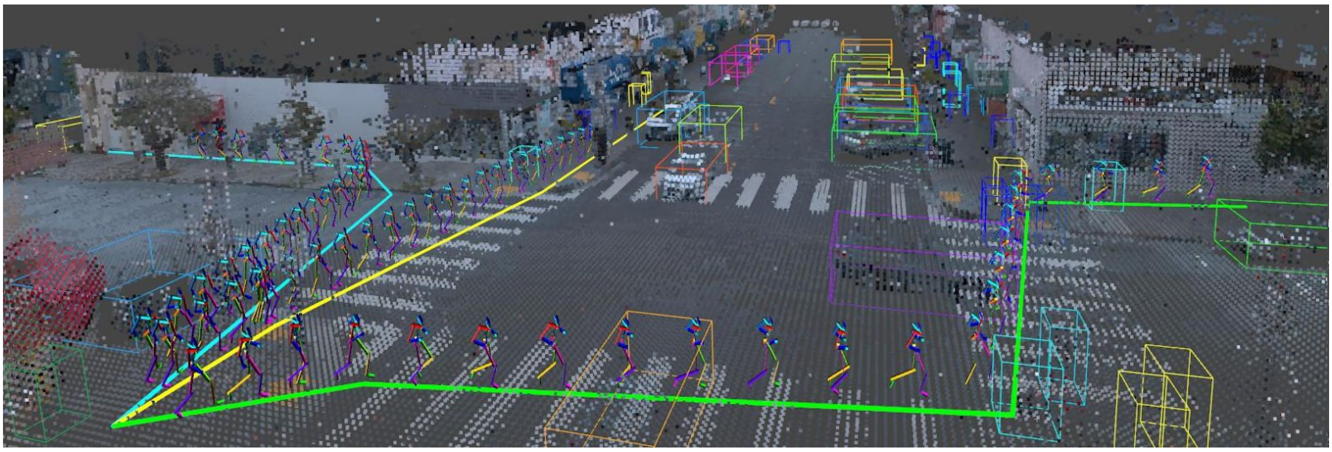
Reference

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 .



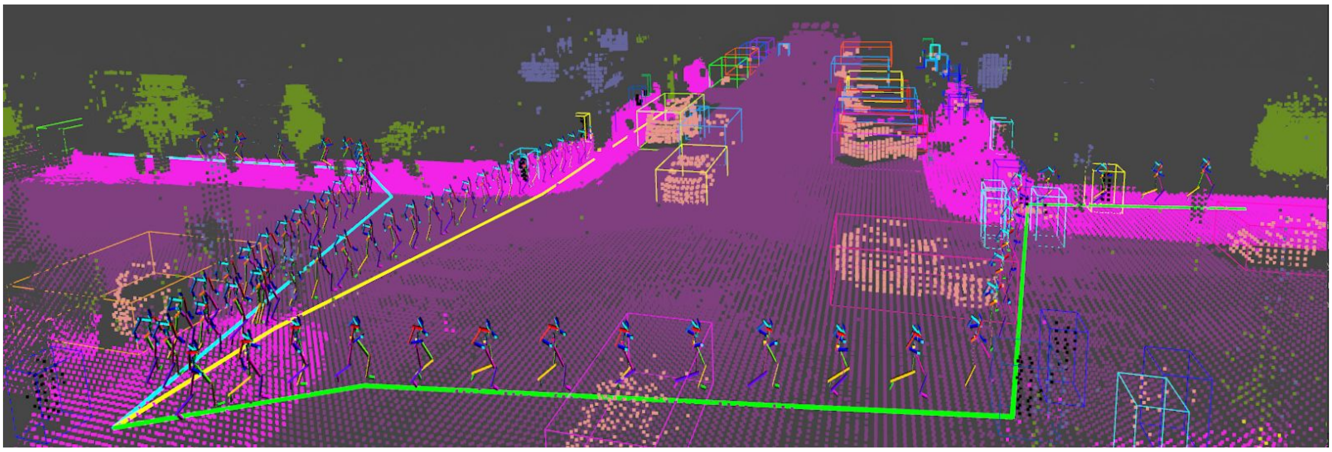


Fig T3.1.1 Pedestrian trajectories and poses generated by our agent on a Waymo scene. RGB and semantic point clouds of the scene are shown in the top and bottom images, respectively. A local neighborhood of these point clouds are observed by the agent. Coloured lines on the ground show different trajectories 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.

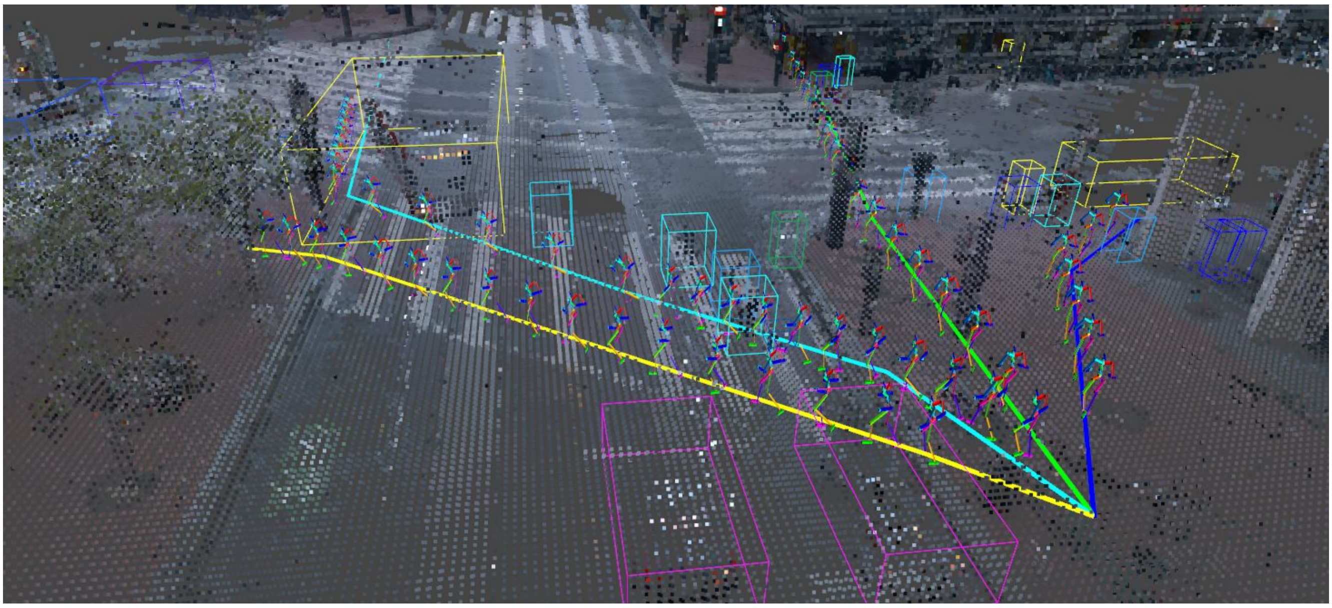


Fig. T3.1.2 SPL agent trajectories on the Waymo dataset, showing the pedestrian taking a number of different paths depending on how the agent history is initialized. Cars and other pedestrians are indicated with 3d bounding boxes. pedestrian trajectories. It should be noted that the collision-aware SPL 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.

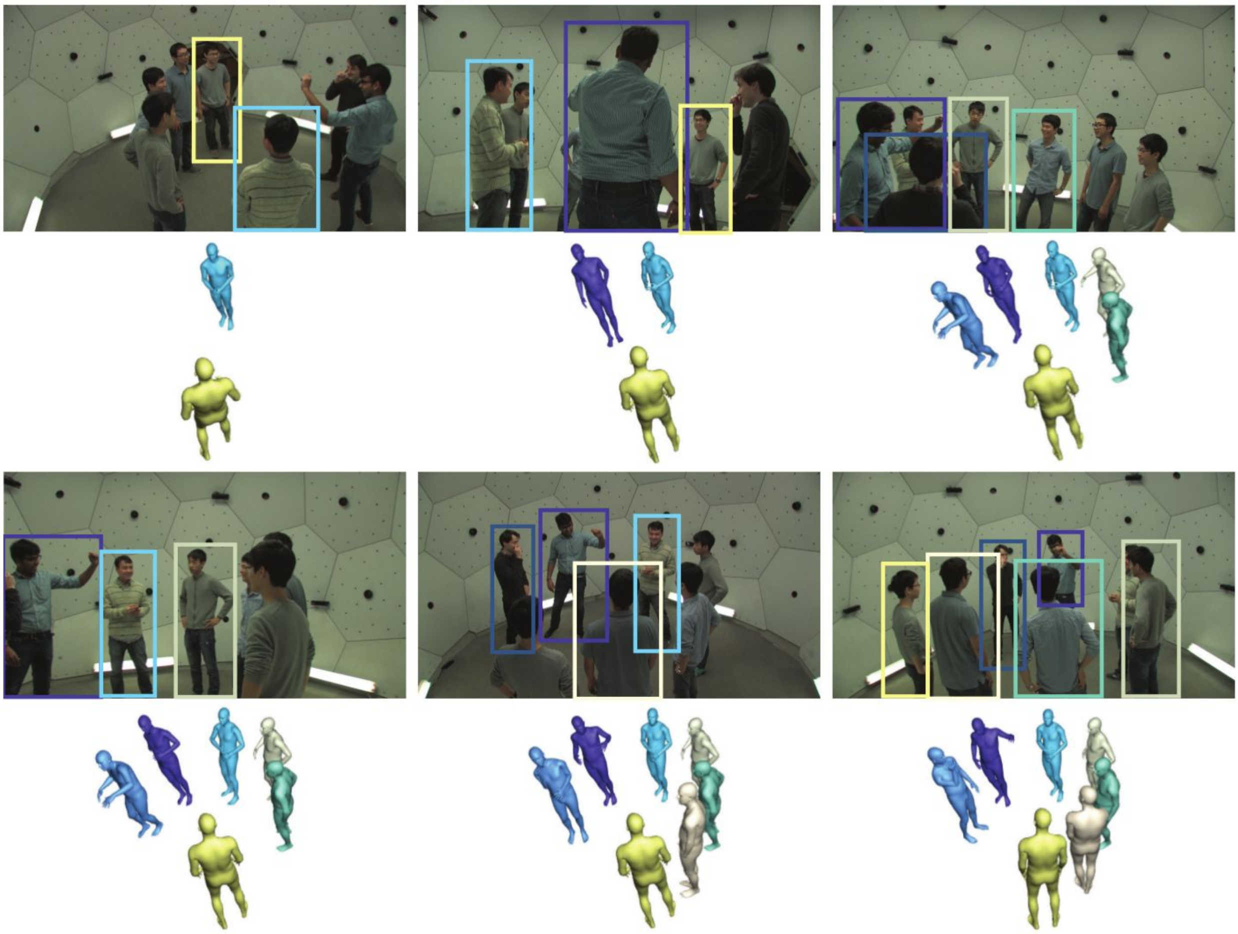


Fig. T3.1.3 Visualization of how Pose-DRL performs multi-target pose estimation for an Ultimatum test scene. In this example the agent sees six viewpoints prior to automatically continuing to the next active-view. The mean error decreases from 358.9 to 114.6 mm/joint. Only two people are detected in 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.

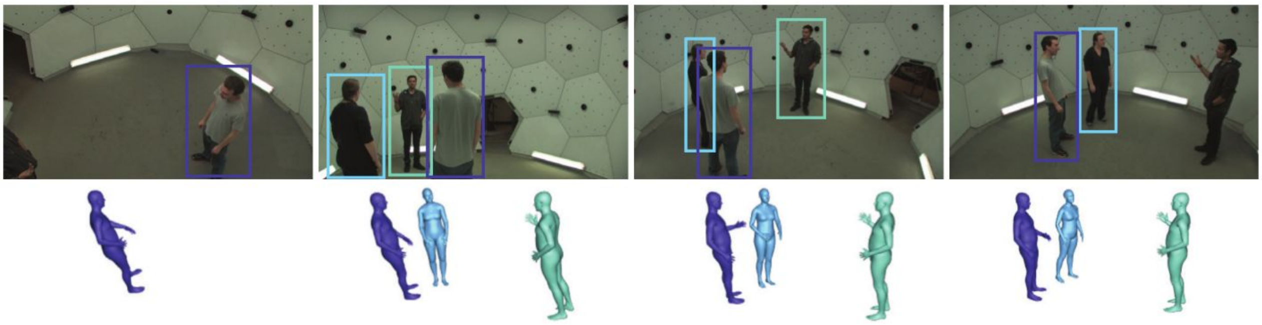


Fig. T3.1.4 Visualization of how Pose-DRL performs multi-target pose estimation for an Ultimatum validation scene. The agent chooses four viewpoints prior to automatically continuing to the next active-view. The mean error decreases from 334.8 to 100.9 mm/joint. Only one of the persons is visible in 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.

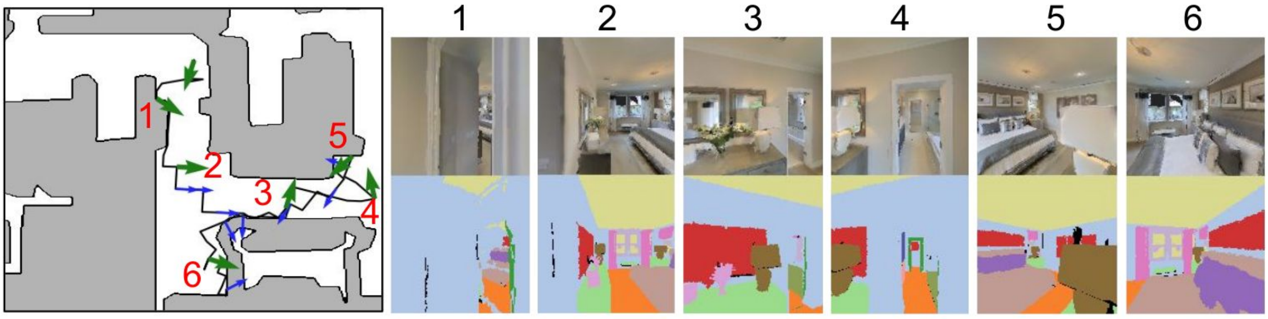


Fig T3.2.1 The first six requested annotations by the RL-agent in a room from the test set. Left: Map showing the agent’s trajectory and the six first requested annotations (green arrows). The initially given annotation is not indicated with a number. Blue arrows indicate Collect actions. Right: For 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.

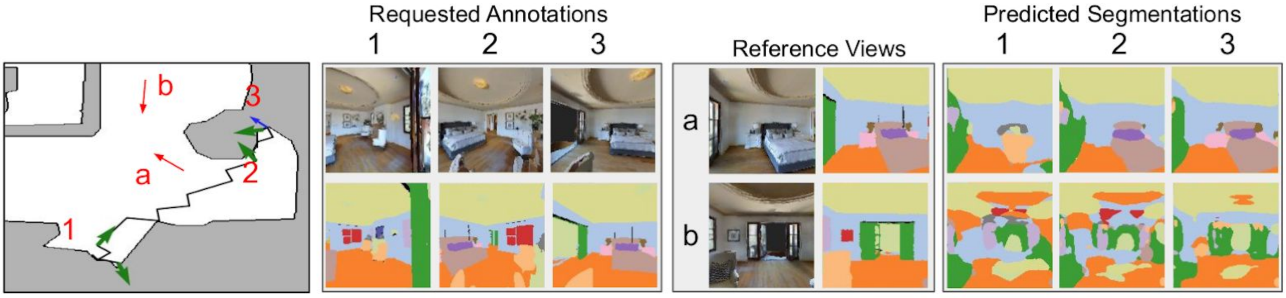


Fig T3.2.2 Example of the RL-agent's viewpoint selection and how its perception improves over time. We show results of two reference views after the first three annotations of the RL-agent. Left: Agent's movement path is drawn in black on the map. The annotations (green arrows) are numbered 1 - 3, and 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)

Existing state-of-the-art estimation systems can detect 2d poses of multiple people in images quite reliably. In contrast, 3d pose estimation from a single image is ill-posed due to occlusion and depth ambiguities. Assuming access to multiple cameras, or given an active system able to position itself to observe the scene from multiple viewpoints, reconstructing 3d pose from 2d measurements becomes well-posed within the framework of standard multi-view geometry. Less clear is what is an informative set of viewpoints for accurate 3d reconstruction, particularly in complex scenes, where people are occluded by others or by scene objects. In order to address the view selection problem in a principled way, we here introduce ACTOR, an active triangulation agent for 3d human pose reconstruction. Our fully trainable agent consists of a 2d pose estimation network (any of which would work) and a deep reinforcement learning-based policy for camera viewpoint selection. The policy predicts observation viewpoints, the number of which varies adaptively depending on scene content, and the associated images are fed to an underlying pose estimator. Importantly, training the policy requires no annotations - given a 2d pose estimator, ACTOR is trained in a self-supervised manner. In extensive evaluations on complex multi-people scenes filmed in a Panoptic dome, under multiple viewpoints, we compare our active triangulation agent to strong multi-view baselines, and show that ACTOR produces significantly more accurate 3d pose reconstructions. We also provide a proof-of-concept experiment indicating the potential of connecting our view selection policy to a physical drone observer. For qualitative results, please have a look at fig. T3.2.3 below.

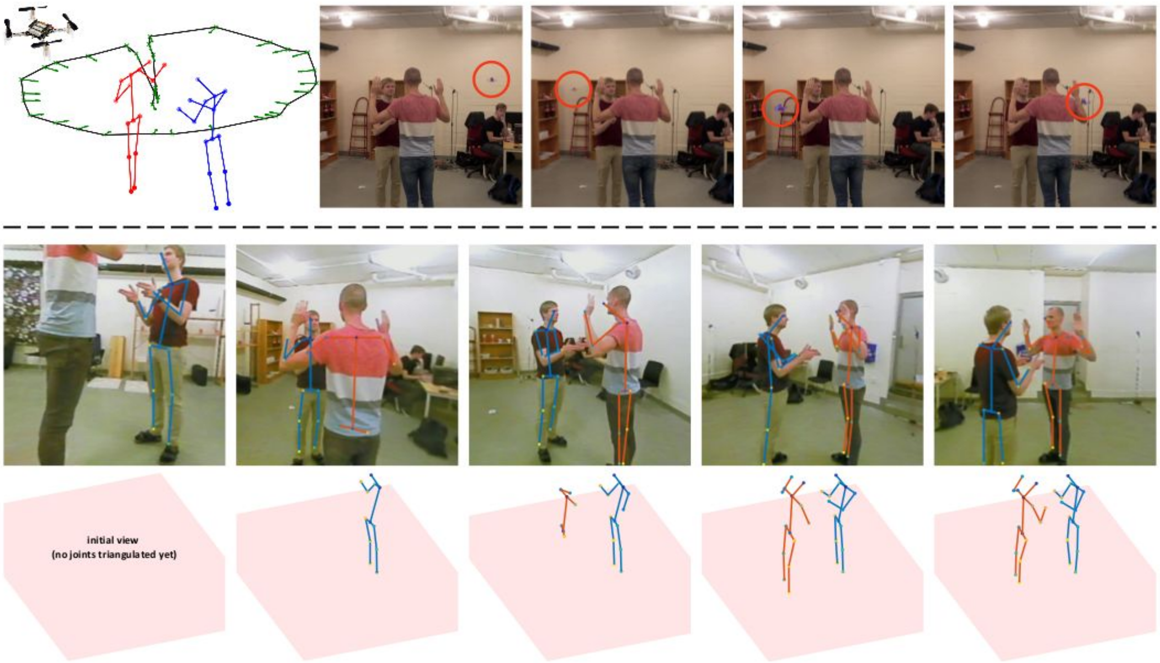


Fig. T3.2.3 From domes to drones. Proof-of-concept experiment illustrating that ACTOR can be connected to an active drone observer to reconstruct 3d poses from informative viewpoints. Above the dashed line to the left we show the drone's loop (the sharp peak is due to take-off and landing), with sampled camera locations as green arrows. We also show the 3d pose reconstructions obtained by triangulating from all 33 sampled camera locations. The 9-by-9 cm Crazyflie drone used is shown in the very top left corner; it can be used safely due to its small size and weight. Sample locations of the drone are also shown above the line (drone locations are highlighted with red circles in images). Below the line we show views seen by ACTOR and aggregated 3d pose reconstructions. After observing 5 viewpoints, the two bodies are fully 3d reconstructed, with an average 2d reprojection error of 11.5 pixels (averaged over all 33 cameras), significantly better than the exhaustively triangulated reconstructions to the left, with an average reprojection error of 35.4 pixels.

Reference

A. Pirinen, E. Gärtner, C. Sminchisescu. "Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstruction", Advances in Neural Information Processing Systems, 2019

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.

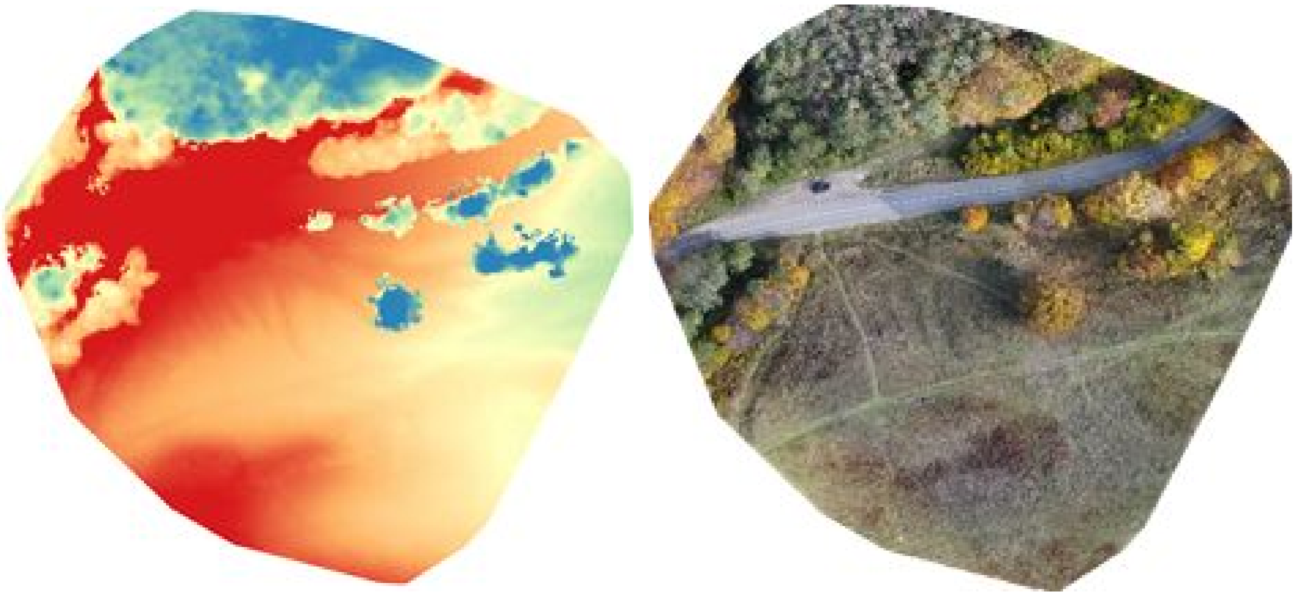
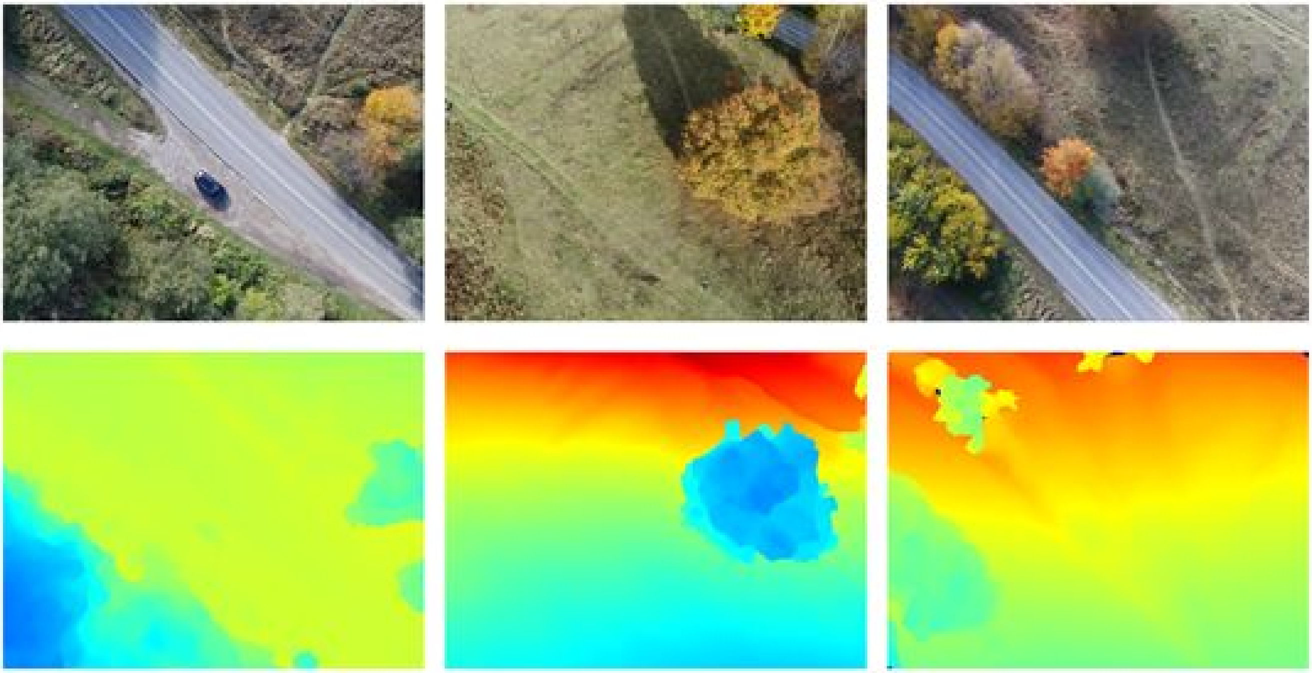


Fig. T4.1.1 Digital Surface Model (DSM) coloured based on elevation and Ortophoto resulting from the mapping process for one of the areas in the dataset

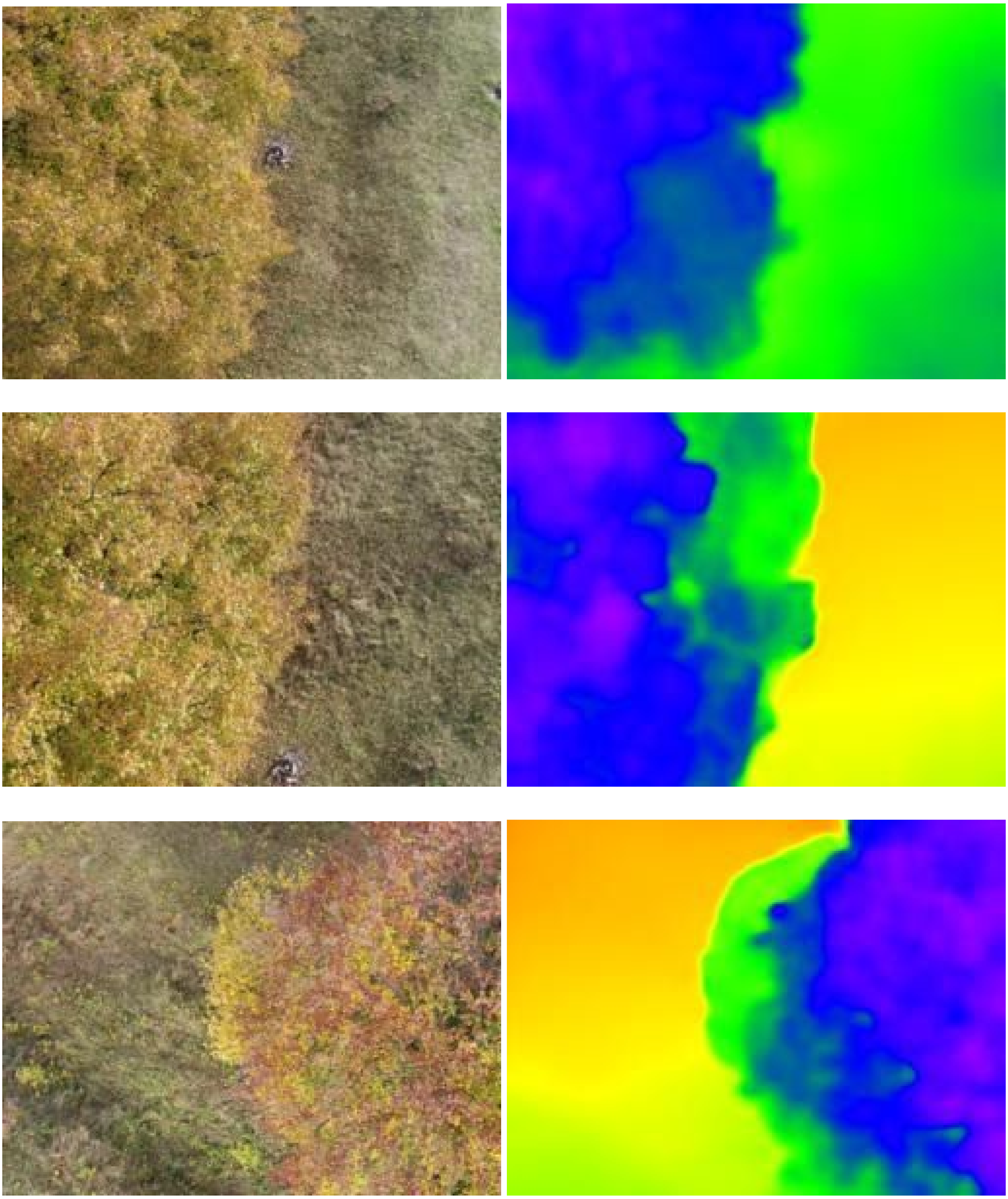


$$R = \begin{pmatrix} -0.784726 & 0.619606 & 0.619606 \\ 0.619813 & 0.784657 & -0.011999 \\ 0.005982 & -0.020015 & -0.999781 \end{pmatrix} \qquad R = \begin{pmatrix} 0.789483 & -0.544615 & 0.283037 \\ -0.613769 & -0.699177 & 0.366657 \\ -0.001793 & -0.463190 & -0.886257 \end{pmatrix} \qquad R = \begin{pmatrix} -0.789156 & 0.535076 & -0.301538 \\ 0.614149 & 0.693240 & -0.377143 \\ 0.007237 & -0.482815 & -0.875692 \end{pmatrix}$$

$$T = \begin{pmatrix} 695031.222 & 5174157.870 & 798.563 \end{pmatrix} \qquad T = \begin{pmatrix} 694983.765 & 5174094.032 & 798.659 \end{pmatrix} \qquad T = \begin{pmatrix} 695035.279 & 5174139.163 & 798.611 \end{pmatrix}$$

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.



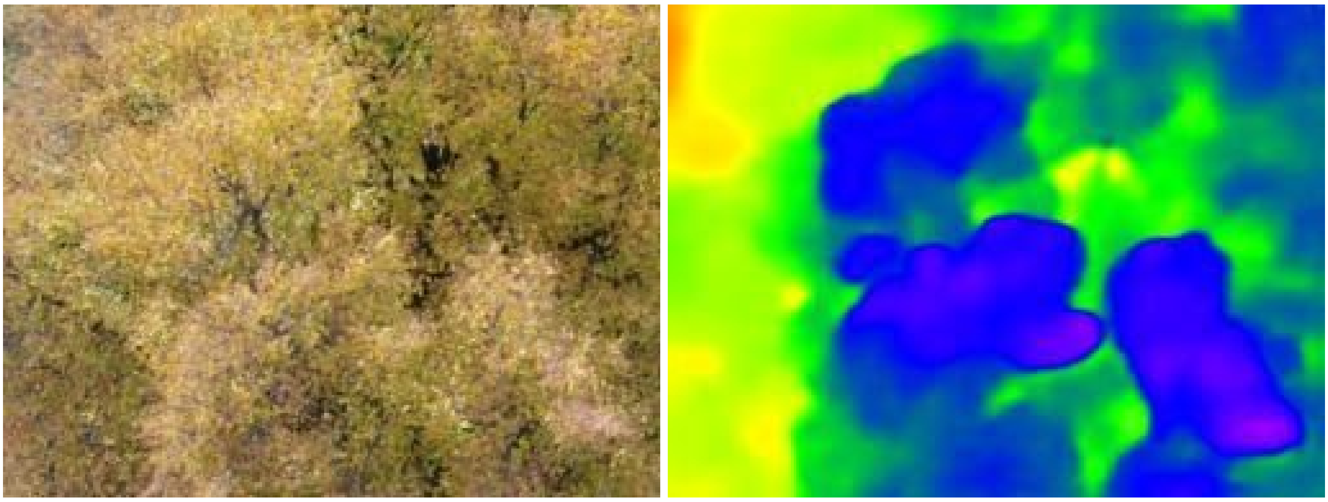


Fig. T4.1.3 Results of a *self-supervised depth and ego-motion estimation solution adapted for aerial images.*

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.

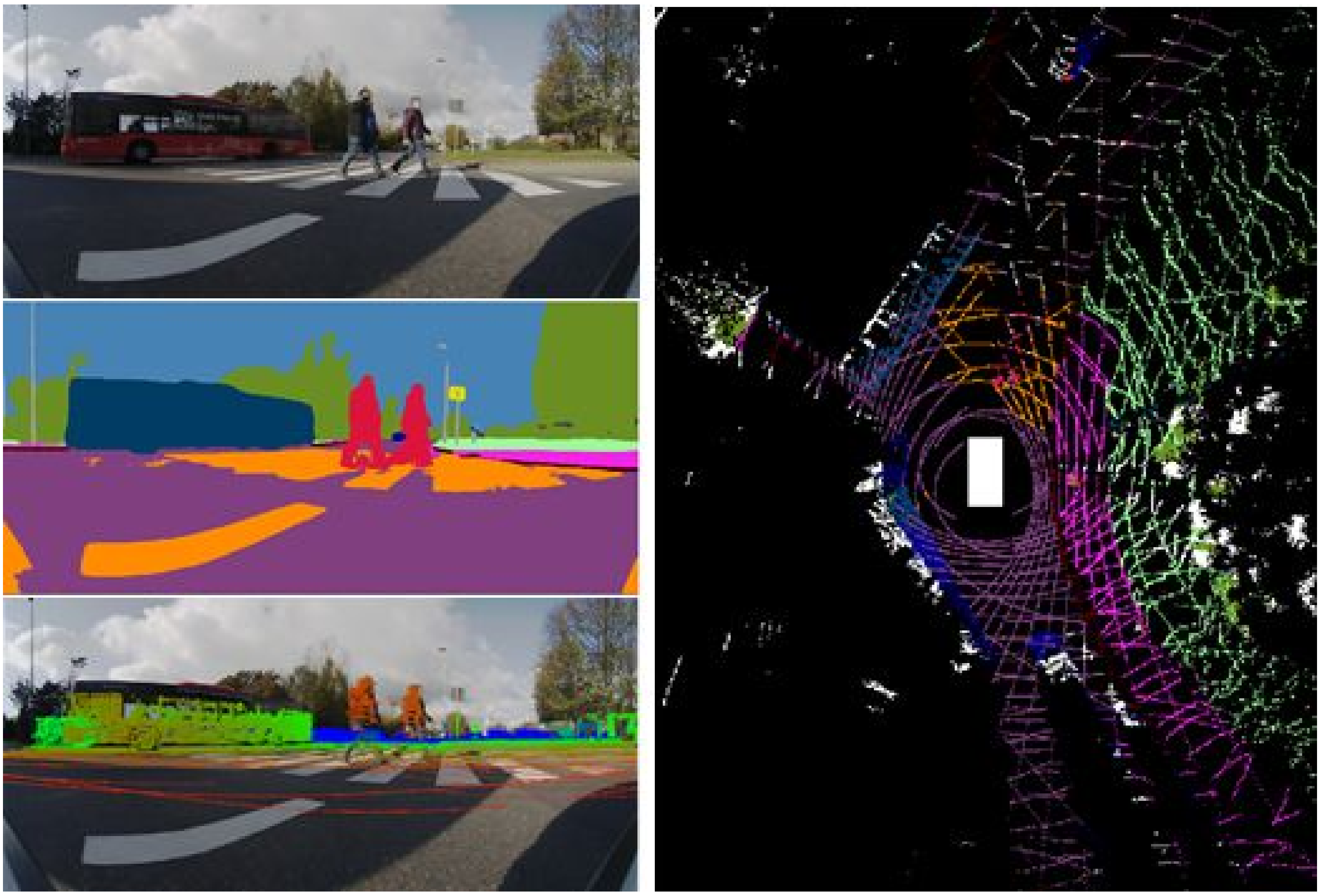


Fig. T4.1.4 Results from the paper “**Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion**”. Low level fusion concept and Spatio-Temporal and Appearance Based Representation (STAR)

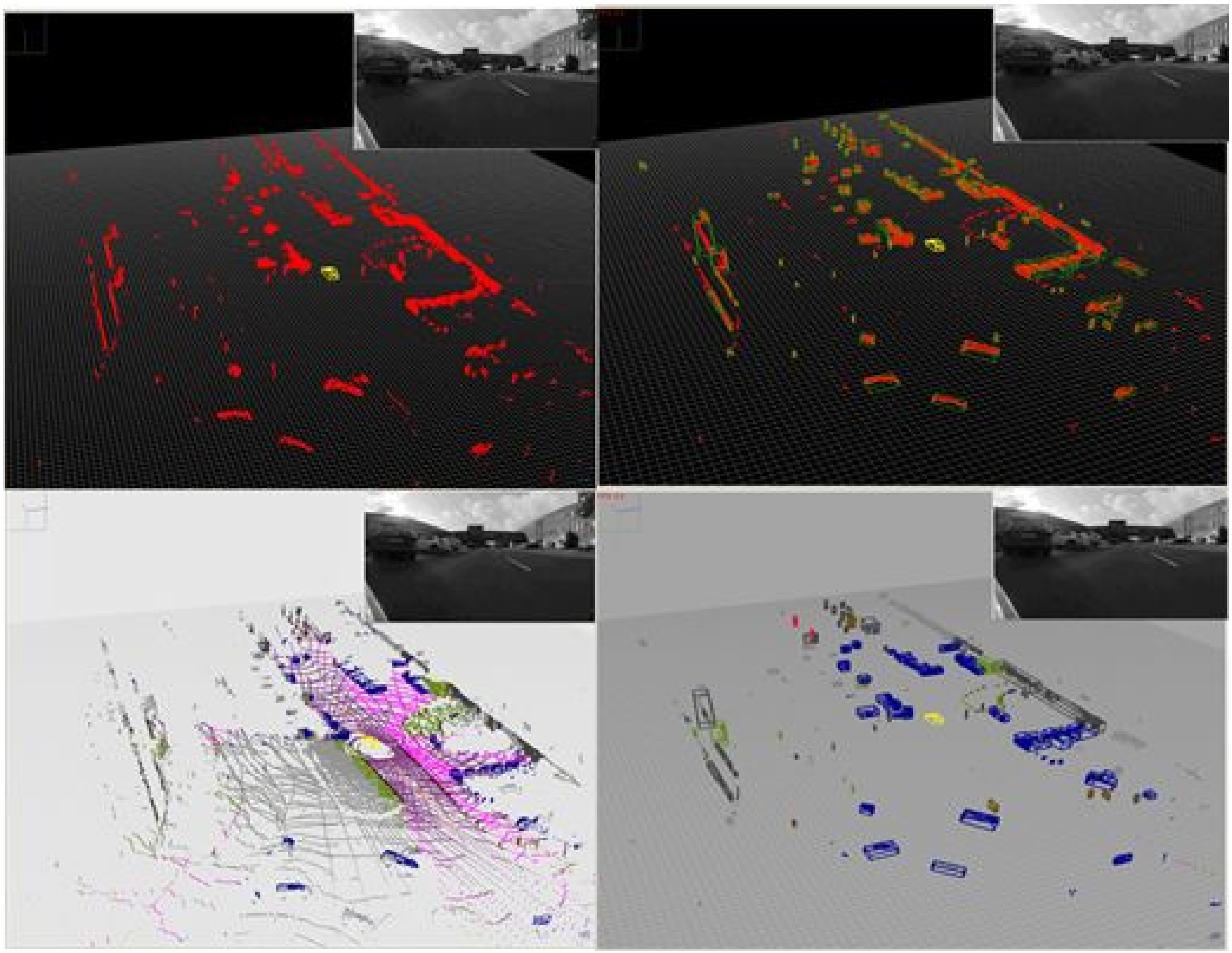


Fig. T4.1.5 Results from the paper “Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion”. Enhanced perception based on the Spatio-Temporal and Appearance Based Representation (STAR)

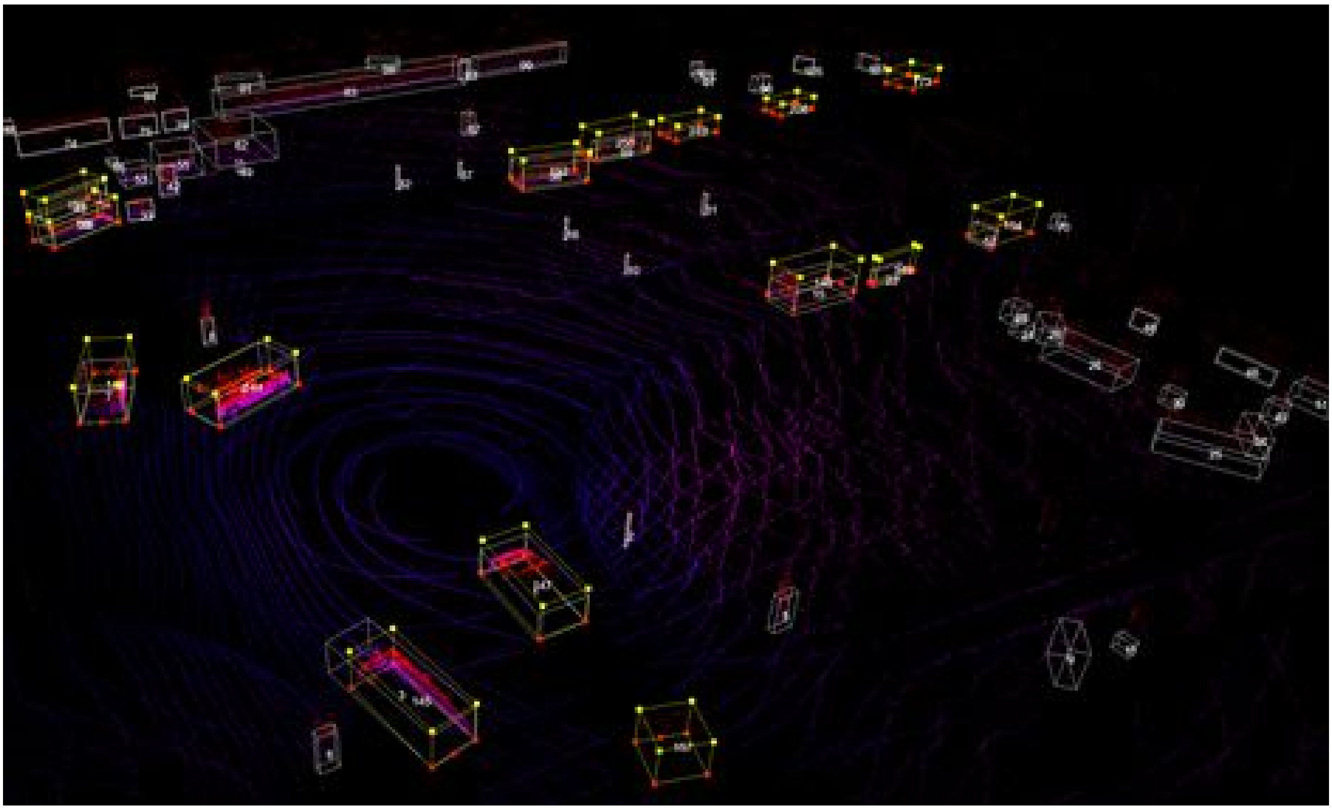


Fig. T4.1.6 Results from the paper “Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion”, Evaluation of detection, classification and localization by comparing 3D perception results with the 3D annotation of the acquired multimodal point cloud.

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

3. **Progress beyond the state of the art and expected results until the end of the project**

A1. Deep 3D Reconstruction Contributions Beyond State-Of-The-Art

Task 1.1: Deep Learning of Graph Matching under Global Constraints

In a best paper honorable mention awarded paper at CVPR 2018, "**Deep Learning of Graph Matching**", we showed how to combine graph matching with deep learning formulations. We propose to build models where the graphs are defined over unary node neighborhoods and pair-wise structures computed based on learned feature hierarchies. We formulate a complete model to learn the feature hierarchies so that graph matching works best: the feature learning and the graph matching model are refined in a single deep architecture that is optimized jointly for consistent results. Methodologically, our contributions are associated to the construction of the different matrix layers of the computation graph, obtaining analytic derivatives all the way from the loss function down to the feature layers in the framework of matrix backpropagation, the emphasis on computational efficiency for backward passes, as well as a voting based loss function. The proposed model applies generally, not just for matching different images of a category, taken in different scenes (its primary design), but also to different images of the same scene, or from a video.

Task 1.2: Deep Structured Geometric Models with Semantics

In a paper published at NIPS 2018, "**Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images**", we proposed a novel, feedforward deep network, supporting different supervision regimes, that predicts the 3d pose and shape of multiple people in monocular images. We formulated and integrated the difficult problem of localizing and grouping people into the network, as a binary linear integer program, and solved it globally and optimally under kinematic problem domain constraints and based on learned scoring functions for body parts that combine 2d and 3d information for accurate reasoning. State-of-the-art results on Human3.6M and Panoptic illustrate the feasibility of the proposed approach. Qualitative examples show that our model can reliably estimate the 3d properties of multiple people in natural scenes, with occlusion, partial views, and complex backgrounds.

With "**HUman Synthesis and Scene Compositing**" (AAAI 2020) we presented a framework (HUSC) that is able to realistically synthesize a photograph of a person, in any given pose and shape, and blend it veridically with a new scene, while obeying 3d geometry and appearance statistics. By operating in the 3d scene space rather than image space, except for late global image adaptation stages, and by taking into account scene semantics, we are able to realistically place the human impostor on support surfaces, handle scene scales and model occlusion. Moreover, by working with parametric 3d human models and dense geometric correspondences, we can better control and localize the appearance transfer process during synthesis. To summarize, our contributions are as follows: (a) a realistic human appearance translation task, with state-of-the-art results (the model produces pleasing qualitative results and obtains superior quantitative results on the DeepFashion dataset), (b) a realistic data augmentation procedure, which allows for the synthesis of complex scenes containing humans, with available pseudo-ground-truth labels such as: pose, shape, segmentation and depth. As a result, the method is practically applicable for photo editing, fashion virtual try-on, or for realistic data augmentation used for training large scale 3d human sensing models.

In "**Three-dimensional Reconstruction of Human Interactions**" (CVPR 2020) we have argued that progress in human sensing and scene understanding would eventually require the detailed 3d reconstruction of human interactions where contact plays a major role, not only for veridical estimates, but in order to ultimately understand fine-grained actions, behavior and intent. One of the main contributions of the paper is a graded modeling framework for Interaction Signature Prediction (ISP) based on contact detection and 3d correspondence estimation over model surface regions at different levels of detail, with subsequent 3d reconstruction under losses that integrate contact and surface normal alignment constraints. Specifically, we propose a first set of methodological elements to address the reconstruction of interacting humans, in a principled manner, by relying on recognition, segmentation, mapping, and 3d reconstruction. Thus, we break down the problem of producing veridical 3d reconstructions of interacting humans into (a) contact detection, (b) binary segmentation of contact regions on the corresponding surfaces associated to the interacting people; (c) contact signature prediction to produce estimates of the potential many-to-many correspondence map between regions in contact; and (d) 3d reconstruction under augmented losses built using additional surface contact constraints given a contact signature. To prove the value of our proposed methodology we have undertaken a major effort to collect 3d ground truth data of humans involved in interactions (CHI3D, 631 sequences containing 2,525 contact events, 728,664 ground truth poses), as well as image annotations in the wild (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 regions). We have evaluated all components in detail, showing their relevance towards accurate 3d reconstruction of human contact. Models and data are made available for research at http://vision.imar.ro/ci3d.

Further down that line of thought, to overcome some of the shortcomings of existing, self-contact agnostic, 3d reconstruction methods, we proposed in "**Learning Complex 3D Human Self-Contact**" (AAAI 2021) to represent self-contact explicitly by learning to predict the image location of contact in order to assist the detection of body regions in self-contact, as well as their signature, defined as the correspondences between regions on the surface of a human body model that touch. To train models and for large-scale quantitative evaluation, we collected and annotated two large scale datasets containing images of people in self-contact. HumanSC3D is an accurate 3d motion capture dataset containing 1,032 sequences with 5,058 contact events and 1,246,487 ground truth 3d poses synchronized with images captured from multiple views. We also collect FlickrSC3D, a dataset of 3,969 images, containing 25,297 annotations of body part region pairs in contact, defined on a 3d human surface model, together with their self-contact localisation in the image. The main contributions of the paper are the following: (a) a first principled model to detect self-contact body regions and their signature by using a novel deep neural network SCP, assisted by an intermediate self-contact image localisation (branch) predictor, leveraged both in training, for local feature selection, and in testing, by enforcing consistency with the estimated 3d contact signature, (b) novel, task-specific, large scale, valuable community datasets capturing people in self-contact, together with dense annotations of a 3d body model to capture the surface regions in contact, as well as image annotations associated with the observed points of contact, (c) quantitative and qualitative demonstration of metrically more accurate and perceptually veridical 3d reconstructions based on self-contact signatures, (d) a foundation for a large class of applications that would benefit from accurate 3d self-contact representations, such as, health monitoring of possible infections when hands touch parts of the face (mouth, nose, eyes) in hospitals or during a pandemic, or subtle behavioral understanding of gestures for robot-assisted therapy of children with autism, to name just a few.

Monocular depth estimation is generally seen as an ill-posed problem since an infinity of 3D scenes can be generated from a single image [19][20][21]. MDE methods deal with this lack of constraints in multiple ways. A first category of algorithms uses stereo information [21] to generate a proper ground truth and to guide the learning process. Other algorithms use guiding information such as semantic segmentation, planes or surface normal to optimize the process. Other algorithms [19][20] use CNNs to perform feature extraction at multiple scales, which are then cleverly combined such that a scene object ordering is established. Thus, both local object composition and relations and global features are captured and then from these relations a depth map can be regressed [19], classified or orderly regressed [20] with respect to a ground truth. DORN [20] belongs to the order regression category and it proves that this approach generates the top results on most of the benchmarks [18].

One of the main problems with supervised MDE approaches is that they still do not ensure robustness when compared to stereo or structure from motion than use geometric constraints to derive the depth. In the article "**SGM-MDE: Semi-global optimization for classification-based monocular depth estimation**" [23] published at the **International Conference on Intelligent Robots and Systems (IEEE IROS) 2020**, **we proposed a novel method to cope with the lack of geometric constraints to monocular depth estimation (MDE)**. **The method approaches the task by initially mathematically transforming the feature vectors from the last layer inside a MDE CNN such that a 3D stereo-like cost volume is generated. Then the semi-global stereo optimization [16] is adapted to the aforementioned volume, further introducing constraints by the global consistency ensured by SGM.** The method can be applied to any classification-based MDE, experiments proving that our technique increases the accuracy and the robustness for any such methods, being also usable for real-time applications. The method can be used to bridge the gap between geometrically-constrained depth perception methods such as stereo reconstruction or structure from motion and single-camera depth estimation, increasing the reliability of the methods in the MDE category. This increased reliability is especially important in the case of aerial environments, which generally lack structure and lead to low robustness.

Another problem generally studied in the context of depth estimation is dealing with objects at large distances [22]. The issue is generally studied in the context of stereo reconstruction, but all camera-based methods suffer from the lack of reliability for long-distance points. In the article "**A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithms**" [24] published at the **2020 IEEE Intelligent Vehicles Symposium (IV)** we introduced a unified method for improving the long-range accuracy of multiple types of camera-based depth estimation algorithms. Towards improving the capabilities of long-range stereo and monocular depth estimation methods, the article initially introduces a taxonomy to categorize all types of camera-based depth perception methods with respect to their long-range capabilities. **A correction method is then introduced, which initially extracts valuable information from neighboring feature vectors and then it statistically learns how to interpolate a fractional depth value from them**. The learning mechanism is based on a stochastic optimization method which properly corrects the depth. The method is shown to work for both stereo and monocular depth perception algorithms that output a depth in discrete setting (most suitable for real-time applications). The article shows that this method improves the precision for such algorithms for objects at large distances without affecting the near-range accuracy. The method requires only several additional operations preserving the real-time capabilities of the underlying algorithms. This novel approach creates an opportunity for generating a 3D representation in new scenarios, extending the set of camera-based applications for both stereo and monocular perception. Since drone (aerial) environments generally assume a larger range of depth measurement, ensuring accuracy for objects at long distances is a key aspect of a reliable depth estimation algorithm.

Our work with respect to supervised monocular depth estimation improvement is continued in the article "**Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception**" [25] submitted in 2020 and accepted for publication in **IEEE Transactions on Geoscience and Remote Sensing** in 2021. In this work we introduced a novel approach for monocular depth estimation, capable to work on complex aerial images, captured from a medium distance, in a variety of scenarios. The method initially proposes an original CNN, particularly adapted to such scenarios. **The novel CNN contains a Darknet-based feature extractor, that is both lightweight and properly built to capture aerial information. Furthermore, our method introduces a new scene understanding module that captures features at multiple scales and it combines them with an image encoder**. **Next, we introduce a new loss that combines the benefits given by**

ordinal regression (that produces very good results for smooth areas) with classification (that better accounts for isolated objects). The most significant contribution in this article is 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 in the aerial image category (a reduction in average error of at least 2 meters with respect to the baseline [20] – the top approach on Kitti benchmark), while the additional refinement further improves the accuracy with only a few additional computational resources. The method is applied on both synthetic and real-life scenarios, providing very accurate results in both field and forest-like environments. This article proved that supervised deep learning can be effectively used for the first step in producing a 3D representation of the environment from aerial perspectives: generating an initial depth map.

Another solution for improving the accuracy and robustness of depth estimation by fusing the geometric models with semantics was already done by studying the relation between geometric and semantic features in context of two-camera (stereo)-based depth estimation. In the article "**Real-time Semantic Segmentation-based Stereo Reconstruction**" [26], published in **IEEE Transactions on Intelligent Transportation Systems** in 2020, we proposed a novel real-time stereo reconstruction solution that accounts for both geometry and semantic in producing an accurate depth map. While traditional stereo methods [15][16] construct well-built geometrical blocks to find the correct pixel correspondences between two images, they generally do not benefit from scene information captured through machine learning so their accuracy performance is limited. Pure learning-based solutions [17], on the other hand, perform extremely well on state of the art benchmarks [18], but they suffer in both terms of robustness and computational performance. Therefore, we propose a method that benefits from both categories of algorithms, **introducing a novel approach to generate a depth map of a scene by aiding the stereo reconstruction geometrical steps with semantic information given by semantic features**. Initially, a semantic map of the scene is generated by using a convolutional neural network. Then, each step of the stereo pipeline: cost computation, aggregation, optimization and refinement is tailored to incorporate scene information obtained from the semantic map and thus to enhance the results. While more classic machine learning-based solutions (based on genetic algorithms) are introduced for the first three steps, a novel real-time CNN is proposed for the refinement. The CNN simultaneously extracts features from the RGB image, the uncorrected depth map and the semantic segmentation and it cleverly combines them towards an improved depth map, using regression as the learning mechanism. Results show that the new method produces the best real-time stereo reconstruction results on the Kitti [18] stereo benchmark. This approach is really important since it clearly shows the benefit of using high-level scene information (provided through the semantic map) for low-level vision tasks required for depth perception. We plan to extend this work by replacing the stereo-based solution with a MDE approach, in which the semantic information will also be used.

A2. Visual Recognition and Localization Contributions Beyond State-Of-The-Art

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

Environment perception is key to many robotics applications. A perception system should provide semantic, geometric and motion information about the environment, which is subsequently used by other modules in the system such as scene understanding and motion planning. The requirements of a perception system from an applicative point of view represent high accuracy and low computational time. We tackled the task of semantic perception of the environment by developing novel image panoptic segmentation algorithms.

Panoptic segmentation enables a complete semantic understanding of the 2D scene by providing pixel-level classification and instance identifiers for dynamic objects.

First solutions on semantic [1] and instance segmentation [2] imply having two separate networks for each task. However, in real-time applications, running two networks is unfeasible due to the high computational costs.

In this context, we introduce in the paper "**Fusion Scheme for Semantic and Instance-level Segmentation**", published at the **Intelligent Transportation Systems Conference 2018**, **a multi-task network for instance and semantic segmentation** [3]. Moreover, **we introduce a fusion scheme in order to provide a coherent output**. In order to speed up the computation, we propose reusing the feature extraction part of an object detection network, such as Mask R-CNN [2] and extending it with a semantic head. **Merging the semantic and instance segmentation into one coherent output is not trivial due to overlaps and conflicts, therefore we introduce a novel fusion approach to refine the outputs of the network based on object sub-category class and instance propagation guidance by semantic segmentation for more general classes**. The proposed solution achieves significant improvements in semantic object segmentation and object mask boundaries refinement at low computational costs. We perform extensive experiments on the Cityscapes dataset, consisting of 5000 high resolution 1024 × 2048 images with urban driving scenes. We measure the semantic segmentation mIoU before and after fusion and achieve a 3.1% improvement, from 72.9 mIoU to 76.0 mIoU.

Panoptic segmentation has been formally defined by Kirillov et al. in [4]. The authors provide a baseline solution, consisting of a state-of-the-art semantic segmentation network [1], an instance segmentation network [2] and propose post-processing heuristics to merge the results. Multi-task networks for semantic and instance segmentation have been proposed in [5], where the authors perform a detailed study of the extension of Mask R-CNN with a novel semantic segmentation head. Xiong et al. [6] design a deformable convolution based semantic segmentation and introduce a parameter-free panoptic segmentation head via pixel-wise classification.

In the paper "**Multi-task Network for Panoptic Segmentation in Automated Driving**", published at the **Intelligent Transportation Systems Conference 2019** [7], we propose **an improved segmentation head on top of the Feature Pyramid Network of Mask R-CNN**, which achieves 73.3 mIoU on the Cityscapes dataset. **The second contribution of our work represents the design of a novel panoptic segmentation head end-to-end trainable with the rest of the network, that avoids hand crafted post-processing steps**. We evaluate our model on the Cityscapes dataset. For semantic segmentation, we obtain 73.3 mIoU from the segmentation head, while the panoptic head improves the semantic segmentation to 75.4 mIoU. We also evaluate the panoptic quality and achieve 57.3 PQ.

Mask R-CNN based approaches for panoptic segmentation are accurate but not suitable for real-time processing. Recently, few works try to strike the balance between real-time performance and quality [8], [9].

In the paper "**Real-Time Panoptic Segmentation with Prototype Masks for Automated Driving**" published at the **Intelligent Vehicles Symposium 2020** [10], we propose a fast fully convolutional neural network for panoptic segmentation that can provide an accurate semantic and instance-level representation of the environment in the 2D space. **We design an end-to-end trainable and lightweight network for this task, thus avoiding complicated post-processing steps. Our architecture is based on the single-shot and anchor-free FCOS object detector [11], which formulates the object detection problem as a per-pixel prediction**. **We extend FCOS with a lightweight semantic segmentation decoder, that is used for stuff class mask prediction**. **We design a novel panoptic segmentation network that generates a fixed number of scene prototype masks, which can be assembled into instance masks guided by object proposals**. Panoptic segmentation is obtained via pixel level classification. We perform extensive experiments on the challenging Cityscapes dataset and achieve state-of-the-art results at 76.9% mIoU and 57.3 PQ. Our solution is more accurate than [8][9] and also faster. Moreover, the fully convolutional nature of the network facilitates deployment for robotic applications.

Proposal-free approaches that first perform semantic segmentation and then clustering of things pixels into instances, usually have high inference speed, but early results lack in accuracy. However, Panoptic DeepLab [12] closes the quality gap between bottom-up and proposal-based approaches. DenseBox [13] is a single shot panoptic segmentation network that is designed for real-time inference by clustering 2D bounding boxes that are usually discarded in the non-maxima suppression step.

In the paper "**SAPSNet: A Soft Attention Panoptic Segmentation Network**", which is under review at IEEE **Transactions on Image Processing** [14], we introduce a novel, fast and accurate **single-shot panoptic segmentation network that employs a shared feature extraction backbone and three network heads for object detection, semantic segmentation, instance-level attention masks**. **Guided by object detections, our new instance-level head learns instance specific soft attention masks based on spatial embeddings, that are instance center offsets. By weighting the semantic segments with the instance-specific soft attention masks, the network is able to directly learn the panoptic output**. Our solution is faster than [12] and [13]. On the Cityscapes dataset, we achieve on par results with [12] for panoptic segmentation with 59.9% PQ and more accurate results than [13].

A3. Semantic Optimal Control Contributions Beyond State-Of-The-Art

Task 3.1. Direct and Inverse Optimal control

In "**Semantic Synthesis of Pedestrian Locomotion**" (ACCV 2020), we proposed a semantic pedestrian locomotion (SPL) agent, a hierarchical articulated 3d pedestrian motion generator that conditions its predictions on both the scene semantics and human locomotion dynamics. Our agent first predicts the next trajectory location and then simulates physically plausible human locomotion to that location. The agent explicitly models the interactions with objects, cars and other pedestrians surrounding it. To allow our SPL agent to learn also from states outside the training set, in §2 we pose the trajectory forecasting problem in the framework of reinforcement learning (RL). We extrapolate the learning signal with an optima-preserving reward signal that additionally involves prior knowledge to promote e.g. collision avoidance. We adapt the RL policy sampling process to simultaneously optimize the trajectory forecasting loss and maximize the reward. Moreover, our analysis can be used to adapt any trajectory forecasting model into a robust articulated pedestrian synthesis model. Our contributions are as follows: (a) we propose an articulated 3d pedestrian motion generator that conditions its predictions on both the scene semantics and human locomotion dynamics. The model produces articulated pose skeletons for each step along the trajectory. (b) we propose and execute a novel training paradigm which combines the sample-efficiency of behaviour cloning with the open-ended exploration of the full state space of reinforcement learning. (c) we perform extensive evaluations on Cityscapes, Waymo and CARLA and show that our model matches or outperforms existing approaches in three different settings: i) for pedestrian forecasting; ii) for pedestrian motion generation; and iii) for goal-directed pedestrian motion generation.

In "**Deep Reinforcement Learning for Active Human Pose Estimation**" (AAAI 2021) we presented Pose-DRL, a fully trainable deep reinforcement-learning based active vision model for human pose estimation. The agent has the freedom to move and explore the scene spatially and temporally, by selecting informative views that improve its accuracy. The model learns automatic stopping conditions for each moment in time, and transition functions to the next temporal processing step in video. We showed in extensive experiments – designed around the dense Panoptic multi-camera setup, and for complex scenes with multiple people – that Pose-DRL produces accurate estimates, and that our agent is robust with respect to the underlying pose estimator used. Moreover, the results show that our model learns to select an adaptively selected number of informative views which result in considerably more accurate pose estimates compared to strong multi-view baselines. Practical developments of our methodology would include e.g. real-time intelligent processing of multi-camera video feeds or controlling a drone observer. In the latter case the model would further benefit from being extended to account for physical constraints, e.g. a single camera and limited speed. Our paper is a key step since it presents fundamental methodology required for future applied research.

Task 3.2. Representations and Methods for Efficient Computation

In "**Embodied Visual Active Learning for Semantic Segmentation**" (AAAI 2021) we have explored the embodied visual active learning task for semantic segmentation 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. Developing robust scene understanding by exhaustive approaches may be difficult, as looking everywhere requires an excessive amount of annotation labor. Instead, an agent might rather self-train online and find a way to select the most informative views to annotate when exploring an environment and to make the most out of them. For the particular task of semantic segmentation, the agent should be able to accurately segment all views in the explored area, after exploring the scene. This requires an exploration policy covering different objects from diverse viewpoints and selecting sufficiently many annotations to train the perception model. The agent can also propagate annotations to different nearby viewpoints using optical flow and then self-train. We developed a battery of methods, ranging from pre-specified ones to a fully trainable deep reinforcement learning-based agent, which we evaluate extensively in the photorealistic Matterport3D environment and conclude that the fully learning-based method outperforms comparable non-learnt approaches, both in terms of accuracy and mIoU, while relying on fewer annotations. Our main contributions are: (a) we study the task of embodied visual active learning, where an agent should explore a 3d environment to acquire visual scene understanding by actively selecting views for which to request annotation. The agent then propagates information by moving in the neighborhood of those views and self-trains; (b) in our setup, visual learning and exploration can inform and guide one another since the recognition system is selectively and gradually refined during exploration, instead of being trained at the end of a trajectory on a full set of densely annotated views; (c) we develop a variety of methods, both learnt and prespecified, to tackle our task in the context of semantic segmentation; (d) we perform extensive evaluation in a photorealistic 3d environment and show that a fully learnt method outperforms comparable pre-specified ones.

In "**Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstruction**" (NIPS 2019), we have presented ACTOR, a deep reinforcement learning-based agent to actively reconstruct 3d poses from 2d estimates via triangulation. Training the viewpoint selection policy requires no annotations and only uses an off-the-shelf 2d human pose estimator for self-supervision. We evaluated the model in complex scenarios with multiple interacting people and showed that by intelligently selecting informative views the agent outperforms strong multi-view baselines in both speed and accuracy. We also provided proof-of-concept results which indicate that ACTOR can be used in single-camera settings, e.g. to control a physical drone observer. Our main contributions are: (a) an active triangulation agent for obtaining 3d human pose reconstructions by using (any) 2d pose (human body joints) estimation network and a deep reinforcement learning-based policy for observer (i.e. camera location and pose) prediction, within a fully trainable system; (b) a Panoptic multi-view framework implementation of our methodology (where the scene can be observed in time-freeze, from a dense set of viewpoints, and over time, providing a proxy for an active observer) whose evaluation using Panoptic shows that our system learns to select camera locations that yield more accurate 3d pose reconstructions compared to strong multi-view baselines; (c) a proof-of-concept experiment indicating the potential of connecting ACTOR to a physical drone observer.

Future work

We plan to continue our work regarding monocular depth estimation in two main directions. Firstly, we plan to include additional scene information to our CNN models. Semantic or surface normal information can provide additional cues for estimating a proper depth map and we plan to properly integrate these tasks into our models.

A different research direction is to integrate the MDE generation approach into a self-supervised MDE system. Having an accurate initial depth map will be a key aspect for improving the current self-supervised systems, which require a proper initial estimation of the depth of a scene. We are currently developing a module for generating depth and ego-motion predictions optimized for aerial video that uses multiple geometric constraints in order to provide accurate results, while also tackling the known limitations regarding the output's scaling factor. This can be addressed by enforcing a scale consistency between predictions which then enables us to integrate limited on-board positioning data for recovering metric depth results. We also plan on integrating scene flow information and our advanced segmentation solution such that the system is better equipped to handle cases that break the static scene assumption, a prerequisite when using view synthesis to supervise the system during training.

In the future, we will continue to work on semantic and instance segmentation, with a focus on video processing, which can better exploit the temporal dimension of data. We will design and implement a general video module that can be plugged-in in any encoder-decoder type of network and that will aggregate feature representations from the past frames and fuse them with the current features for improved results. Our video module will be based on the Transformer Network, which is often used in the context of natural language processing and is inspired from retrieval systems. We will develop a memory module that will hold feature representations from the past frames and contain rich historic information. Each entry in the memory will be also associated with a key, that is a compressed embedding of the entry. The keys will be used to select relevant values from the memory with respect to the current frame. We will design a fusion module based on recurrent conv-GRUs that aggregates past features with current ones, in order to provide more meaningful input to the decoder part of the network for increased accuracy. Since creating video semantic and instance annotations is a very time-consuming and expensive process, our networks will be trained in a weakly-supervised regime, where only every k images in a video sequence have annotations.

Starting from the achievements in video semantic and instance segmentation and from the video self supervised scale consistent depth and egomotion estimation a novel semantic SLAM solution will be provided exploiting the benefits of geometric, semantic and temporal information fusion.

We also envision applications for the extraction of tree instances from the environment, in two scenarios: crowns seen from above and trunks seen from lateral, as presented in [31]. Such an application will exploit the semantically enhanced 3D map of the environment, and will include a drone control module to allow automatic navigation and survey of the map. The second application is focused around tracking a person in the scene, with the ability to follow the person while avoiding the collision with other objects. Moreover, when the human sends a signal (using, for example, the arm) it will create a recording of the indicated tree or area.

References

[1] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In CVPR, 2017

[2] K. He, G. Gkioxari, P. Dolla ´r, and R. Girshick. Mask r-cnn. In ICCV, 2017

[3] A. D. Costea, A. Petrovai, S. Nedevschi, ”Fusion Scheme for Semantic and Instance-Level Segmentation”, Proceedings of 2018 IEEE Intelligent Transportation Systems Conference (ITSC), Maui, Hawaii, USA, November 4-7, 2018, pp. 3469-3475

[4] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, Panoptic segmentation. In CVPR 2019

[5] A. Kirillov, K. He, R. Girshick, and P. Dollar, Panoptic Feature Pyramid Networks. In CVPR 2019

[6] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun. Upsnet: A unified panoptic segmentation network. In CVPR 2019

[7] A. Petrovai, S. Nedevschi, „Multi-Task Network for Panoptic Segmentation in Automated Driving”, Proceeding of 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 26-30 October,2019

[8] D.deGeus, P.Meletis, and G.Dubbelman. Fast panoptic segmentation network. arXiv preprint arXiv:1910.03892, 2019.

[9] T.-J. Yang, M. D. Collins, Y. Zhu, J.-J. Hwang, T. Liu, X. Zhang, V. Sze, G. Papandreou, and L.-C. Chen. Deeperlab: Single-shot image parser. arXiv preprint arXiv:1902.05093, 2019.

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

[11] Z. Tian, C. Shen, H. Chen and T. He, "Fcos: Fully convolutional one-stage object detection", Proceedings of the IEEE International Conference on Computer Vision, pp. 9627-9636, 2019.

[12] B. Cheng, M. D. Collins, Y. Zhu, T. Liu, T. S. Huang, H. Adam, and L.-C. Chen. Panoptic-deeplab. In CVPR 2020

[13] R. Hou, J. Li, A. Bhargava, A. Raventos, V. Guizilini, C. Fang, J. Lynch, and A. Gaidon. Real-time panoptic segmentation from dense detections. In CVPR 2020

[14] A. Petrovai, S. Nedevschi, SAPSNet: A Soft Attention Panoptic Segmentation, submitted at IEEE Transactions on Image Processing, 2020.

[15] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms", Int. J. Comput. Vis., vol. 47, no. 1, pp. 7-42, Apr. 2002

[16] Hirschmuller, Heiko. "Stereo processing by semiglobal matching and mutual information." IEEE Transactions on pattern analysis and machine intelligence 30.2 (2007): 328-341.

[17] A. Kendall et al., "End-to-End Learning of Geometry and Context for Deep Stereo Regression," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 66-75, doi: 10.1109/ICCV.2017.17.

[18] Geiger, Andreas, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite." 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012.

[19] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” CoRR, vol. abs/1406.2283, 2014.

[20] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular depth estimation,” in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018

[21] C. Godard, O. M. Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6602–6611, 2016.

[22] P. Pinggera, D. Pfeiffer, U. Franke, and R. Mester, “Know your limits: Accuracy of long range stereoscopic object measurements in practice,” in Computer Vision ECCV 2014, ser. Lecture Notes in Computer

Science, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Springer International Publishing, 2014, vol. 8690, pp. 96–111.

[23] V. -C. Miclea and S. Nedevschi, “SGM-MDE: Semi-global optimization for classification-based monocular depth estimation”, 2020, IEEE International Workshop on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 2020.

[24] V. -C. Miclea and S. Nedevschi, "A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithms," 2020 IEEE Intelligent Vehicles Symposium (IV), Las Vegas, NV, USA, 2020, pp. 1234-1241, doi: 10.1109/IV47402.2020.9304598.

[25] V. -C. Miclea and S. Nedevschi, “Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception”, IEEE Transactions on Geoscience and Remote Sensing, Accepted, to appear in 2021.

[26] V. Miclea and S. Nedevschi, "Real-Time Semantic Segmentation-Based Stereo Reconstruction," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 4, pp. 1514-1524, April 2020, doi: 10.1109/TITS.2019.2913883.

[27] DJI Matrice 210 V2 RTK - https://www.dji.com/matrice-200-series-v2

[28] OpenDroneMap - https://www.opendronemap.org/

[29] OpenSfM, Mapillary - https://github.com/mapillary/OpenSfM

[30] MVE – A Multi-View Reconstruction Environment, Simon Fuhrmann, Fabian Langguth and Michael Goesele In: Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage, Darmstadt, Germany, 2014

[31] Wang, D., “Unsupervised semantic and instance segmentation of forest point clouds”. ISPRS Journal of Photogrammetry and Remote Sensing, 165, pp.86-97, 2020

[32] Shah, S., Dey, D., Lovett, C. and Kapoor, A., Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In *Field and service robotics* (pp. 621-635). Springer, Cham, 2018.

4. Results indicators

Indicators	Description / Name	No.
Articles published/accepted/under evaluation in ISI indexed journals	*Article title/Year/DOI/ISSN or eSSN/Journal/ Authors/ Status (under evaluation/accepted/published)* *A.Pirinen and C. Sminchisescu: Deep reinforcement Learning for Visual Object Detection, submitted to International Journal of Computer Vision, 2021. (Impact factor 2019: 5.698)* *H. Petzka and C. Sminchisescu: Non-attracting Regions of Local Minima in Deep and Wide Neural Networks, accepted in Journal of Machine Learning Research, to appear 2021. (Impact factor 2018: 4.091)* *Real-Time Semantic Segmentation-Based Stereo Reconstruction/ 2020/ 10.1109/TITS.2019.2913883/ 1524-9050/ IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS/ Vlad Miclea, Sergiu Nedevschi (Impact factor 6.319)* *Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception/2020/ accepted IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING/ Vlad Miclea, Sergiu Nedevschi (Impact factor 2020: 5.855)* SAPSNet: A Soft Attention Panoptic Segmentation /2020 /under evaluation *IEEE Transactions on Image Processing* / Andra Petrovai, Sergiu Nedevschi (Impact factor: 9.34) *Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion/2020/ under evaluation IEEE Transactions on Intelligent Transportation Systems/ H. Florea, A. Petrovai, I. Giosan, F. Oniga, R. Varga, S. Nedevschi (Impact factor 6.319)* Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving/*2020/ under evaluation IEEE Transactions on Intelligent Transportation Systems/ Andra Petrovai, Sergiu Nedevschi (Impact factor 6.319)*	*7(4/2/1)*
Articles published/accepted/under evaluation in BDI indexed journals	*Article title/Year/Journal/Authors/Status(under evaluation / accepted/published)*	
Patent applications filed nationally and internationally	*Patent title/Issuing authority/Submission date*	1

	IMAGE PROCESSING METHOD, SYSTEM AND DEVICE *European Patent Office* *Publication: 11.12.2019* *Date of filing: 05.06.2019* *https://data.epo.org/gpi/EP3579198A1-IMAGE-PROCESSING-METHOD-SYSTEM-AND-DEVICE*	
Patents obtained at national and international level	*Patent title/Issuing authority/Issue date*	
Conferences attendance	*Conference name/Type/Title/Year* *A. Zanfir and C. Sminchisescu. "Deep Learning of Graph Matching", Proceedings - 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2018 (best paper award honorable mention) - A* conference* *A. Zanfir, E. Marinoiu (Oneata), M. Zanfir, A. Popa, C. Sminschisescu. "Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images", Proceedings of the Thirty-second Conference on Neural Information Processing Systems, NIPS 2018 - A* conference* *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) - A* conference* *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) - A* conference* *M. Fieraru, M. Zanfir, E. Oneata, A. Popa, V. Olaru, C. Sminchisescu, "Learning Complex 3D Human Self-Contact", Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021) - A* conference* *M. Priisalu, C. Paduraru, A. Pirinen, C. Sminchisescu, "Semantic Synthesis of Pedestrian Locomotion", Proceedings of the Asian Conference on Computer Vision (ACCV), 2020 - B conference* *E. Gärtner, A. Pirinen, C. Sminchisescu. "Deep Reinforcement Learning for Active Human Pose Estimation", Proceedings of the Thirty-Fourth AAAI*	25

	Conference on Artificial Intelligence, 2020 - A conference* *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) - A* conference* *A. Pirinen, E. Gärtner, C. Sminchisescu. "Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstruction", Advances in Neural Information Processing Systems, 2019 - A* conference* *A.D. Costea, A. Petrovai, S. Nedevschi, "Fusion Scheme for Semantic and Instance-Level Segmentation", Proceedings of 2018 IEEE Intelligent Transportation Systems Conference (ITSC), Maui, Hawaii, USA, November 4-7, 2018, pp. 3469-3475.* *V. Miclea, S. Nedevschi, "Real-Time Semantic Segmentation-Based Depth Up Sampling Using Deep Learning", Proceedings of 2018 IEEE Intelligent Vehicles Symposium (IV), Changzhou, China, June 26-30, 2018, pp. 300-306, - B conference* *V. Miclea, S. Nedevschi, L. Miclea, "Real-Time Stereo Reconstruction Failure Detection and Correction Using Deep Learning", Proceedings of 2018 IEEE Intelligent Transportation Systems Conference (ITSC), Maui, Hawaii, USA, November 4-7, 2018, pp. 1095-1102.* *A. Petrovai, S. Nedevschi, "Efficient instance and semantic segmentation for automated driving", Proceeding of 2019 IEEE Intelligent Vehicles Symposium (IV 2019), Paris; France; 9 - 12 June, 2019, pp. 2575-2581, - B conference* *A. Petrovai, S. Nedevschi, „Multi-Task Network for Panoptic Segmentation in Automated Driving", Proceeding of 2019 IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand, 26-30 October,2019* *A. Baraian, S. Nedevschi, „Improved 3D Perception based on Color Monocular Camera for MAV exploiting Image Semantic Segmentation", Proceeding of 2019 IEEE Intelligent Computer Communication and Processing (ICCP), 2019, pp. 295-301*	

	B. C. Z. Blaga, S. Nedevschi, „Semantic Segmentation Learning for Autonomous UAVs using Simulators and Real Data”, Proceeding of 2019 IEEE Intelligent Computer Communication and Processing (ICCP), 2019, pp. 303-310 *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), October 25-29, 2020, Las Vegas, SUA, - A conference* *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), October 19–November 13, 2020, Las Vegas, SUA, - B conference* *A. Petrovai, S. Nedevschi, Real-Time Panoptic Segmentation with Prototype Masks for Automated Driving, Proceedings of 2020 IEEE Intelligent Vehicles Symposium (IV2020), October 19–November 13, 2020, Las Vegas, SUA, - B conference* *R. Beche, S. Nedevschi, Narrowing the semantic gap between real and synthetic data, Proceedings of 2020 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP2020), September 3-5, 2020, Cluj-Napoca, Romania.* *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.* *B. Maxim, S. Nedevschi, Efficient spatio-temporal point convolution, Proceedings of 2020 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP2020), September 3-5, 2020, Cluj-Napoca, Romania.* *S. Baciu, F. Oniga, S. Nedevschi, Semantic 3D Obstacle Detection Using an Enhanced Probabilistic Voxel Octree Representation, Proceedings of 2020 IEEE International Conference on Intelligent Computer Communication and Processing (ICCP2020), September 3-5, 2020, Cluj-Napoca, Romania.* *B. C. Z. Blaga, S. Nedevschi, A Critical Evaluation of Aerial Data Datasets for Semantic Segmentation, Proceedings of 2020 IEEE International Conference on*	

	Intelligent Computer Communication and Processing (ICCP2020), September 3-5, 2020, Cluj-Napoca, Romania.	
Books	*Title/Publishing year/Publishing house/ISBN*	
Book chapters	*Book title/chapter title/Publishing year/Publishing house/ISBN*	
Other results		

5. Equipment use

Two DGX-1 equivalent systems (i.e., architecture compliant) with 8 Tesla V100 GPU boards each have been purchased to help achieve the goals of the project. Both computing systems are instrumental in training the deep network models presented in this report. We use frameworks like Tensor-flow and PyTorch to train the models.

Two DJI Matrice 210 RTK drones have also been purchased in order to collect forest data. Aerial images (videos) of forests as well as ground-level records have been gathered and serve as training data for the deep models currently being developed.

6. Difficulties encountered in implementing the project

A significant hindering has been represented by the acquisition process of the DGX-1 equivalent systems. Not only has the tender taken unusually long (roughy 6 months during 2019) due to extensive procedural checks performed by the national authority for public acquisitions, but, unfortunately, NVIDIA, the maker of the Tesla V100 boards experienced production problems at the beginning of the year 2020, which resulted in an additional delay of roughly 3 months for delivering the boards to the vendor of the servers. Given the delay experienced at the beginning of 2019 by the financial opening of the funding body UEFISCDI and the fact that the project started in July 2018 leaving us with insufficient time to start a tender to purchase the servers at the beginning of the project, we found ourselves in the position of having the computing infrastructure available for project purposes after almost two years from the start of the project, sometime in April 2020.

Additionally, the COVID-19 outbreak in 2020 significantly impeded our endeavours to collect the data with the drones purchased in the project, given the lockdown period and the extensive restriction rules enforced by the authorities.

Date 22.02.2021 **Project manager**
Cristian Sminchisescu