Scientific - Technical Report

(2018 - 2022)

Title:	**Integrated Semantic Visual Perception and Control for Autonomous Systems (SEPCA)**
Project web page:	http://vision.imar.ro/sepca
Coordinator (CO):	„Simion Stoilow" Institute of Mathematics of the Romanian Academy (IMAR)
Project partner 1 (P1):	Technical University of Cluj-Napoca (UTCN)

1. Summary and overall objectives of the project

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.

During the project 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 concerned visual learning and recognition in simple scenarios. To this end, we have 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 driving.

An important goal of the project 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, the report shows the results 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, data has been collected for forest navigation, labelling and analysis. Using 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 are accurately positioned based on DGPS information and enable further visual localization tasks.

We consider that all of the results of the project, including methodologies and models, that we highlight 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 undertaken during the project.

2. Achievements of the project

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 neighbourhood 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 neighbourhoods, 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.

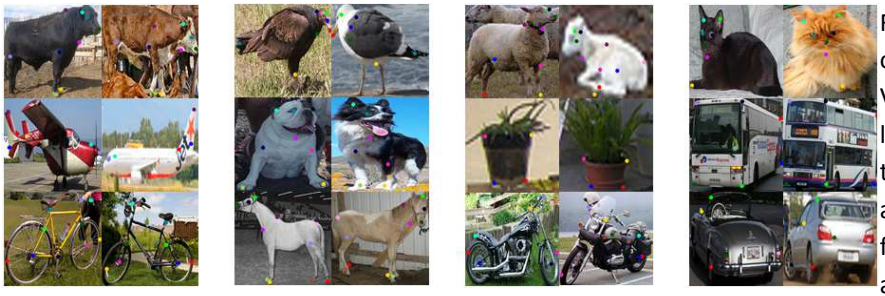


Fig. T1.1.1 Twelve qualitative examples of our best performing network on the PASCAL VOC test-set. For every pair of examples, the left shows the source image and the right the target. Colors identify the computed assignments between points. The method finds matches even under extreme appearance and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

***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.

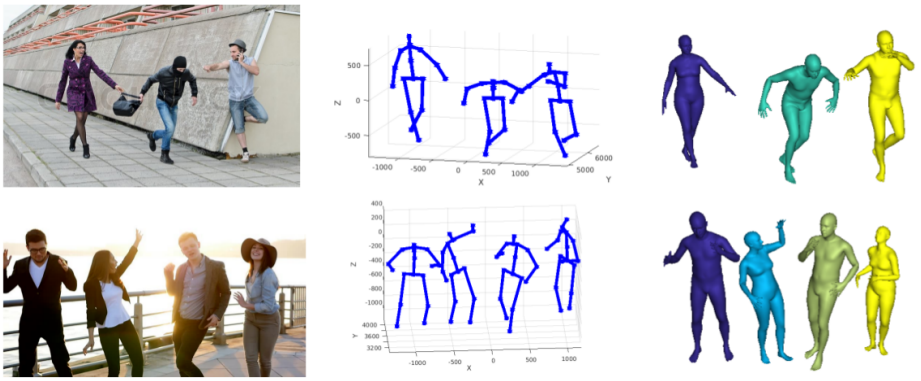


Fig T1.2.1 Human pose and shape reconstruction of multiple people produced by MubyNet illustrate good 3d estimates for distant people, complex poses or occlusion. For global translations, we optimize the Euclidean loss between the 2d joint detections and the projections predicted by our 3d models.

Reference *:* A. Zanfir, E. Marinoiu (Oneata), M. Zanfir, A. Popa, C. Sminchisescu. "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.

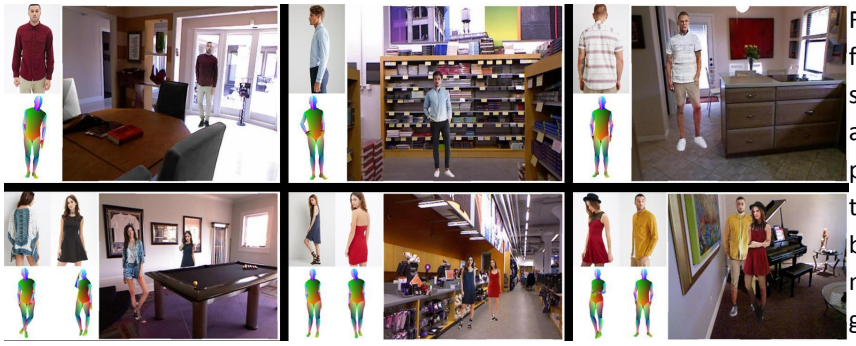


Fig T1.2.2 Sample images generated by our framework. For each example, we show the source image, the target 3d body mesh, and a scene with a geometrically plausible placement of the synthesised person. Note that our framework allows for a positioning behind various objects, and the insertion of multiple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

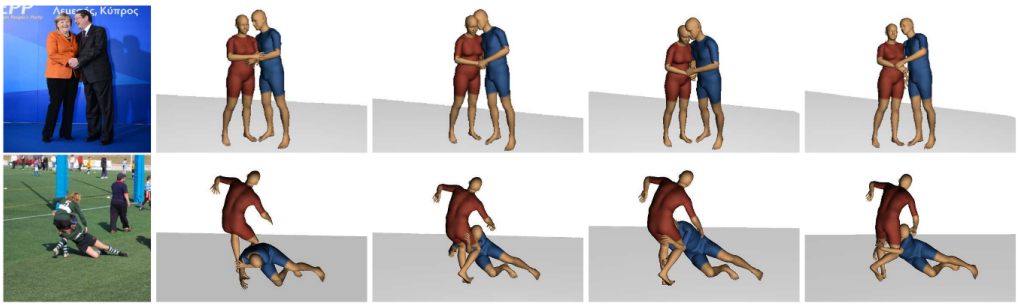


Fig. T1.2.3 3D human pose and shape reconstructions using contact constraints of different granularity. The first column shows the RGB images followed by their reconstructions without contact information (column 2), using contacts based on 37 and 75 regions, respectively (columns 3 & 4), and using facet-based 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 don't focus on body regions in self-contact and thus 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. We developed a model for Self-Contact Prediction (SCP), that estimates the body surface signature of self-contact, 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 enabled by more accurate self-contact models. Examples of qualitative results of our method are shown in fig. T1.2.4.

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)

The three-dimensional reconstruction of multiple interacting humans given a monocular image is crucial for the general task of scene understanding, as capturing the subtleties of interaction is often the very reason for taking a picture. Current 3D human reconstruction methods either treat each person independently, ignoring most of the context, or reconstruct people jointly, but cannot recover interactions correctly when people are in close proximity. In this work, we introduce REMIPS, a model for 3D Reconstruction of Multiple Interacting People under Weak Supervision. REMIPS can reconstruct a variable number of people directly from monocular images. At the core of our methodology stands a novel transformer network that combines unordered person tokens (one for each detected human) with positional-encoded tokens from image features patches. We introduce a novel unified model for self- and interpenetration-collisions based on a mesh approximation computed by applying decimation operators. We rely on self-supervised losses for flexibility and generalisation in-the-wild and incorporate self-contact and interaction-contact losses directly into the learning process. With REMIPS, we report state-of-the-art quantitative results on common benchmarks even in cases where no 3D supervision is used. Additionally, qualitative visual results show that our reconstructions are plausible in terms of pose and shape and coherent for challenging images, collected in-the-wild, where people are often interacting. Qualitative results of the method may be observed in the following picture:

Fig. T1.2.5 3D human pose and shape predictions on the COCO validation set for in-the-wild images. We show the initial image together with an overlaid reconstruction of the meshes as well as a rendering from a different viewpoint which better illustrates the physical consistency of the REMIPS reconstructions

Reference : Mihai Fieraru, Mihai Zanfir, Teodor Alexandru Szente, Eduard Gabriel Bazavan, Vlad Olaru, Cristian Sminchisescu. REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People under Weak Supervision. Proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 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. T1.2.6 depicts semantic segmentation of deforestation areas from aerial images as well as instance level segmentation of trees within the forest.

Fig. T1.2.6 Aerial images with semantic segmentation of the non-forest areas (top row) and forest level images with instance segmentations of the trees and semantic segmentation of the forest (bottom row). 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.

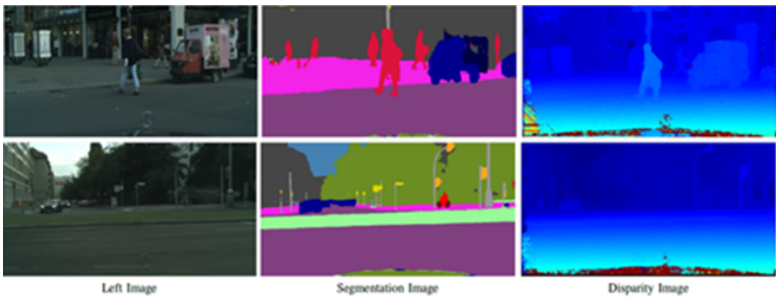


Fig. T1.2.7 The disparity image (from Cityscapes dataset) is obtained by enhancing the stereo pipeline with information from the semantic segmentation.

***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.

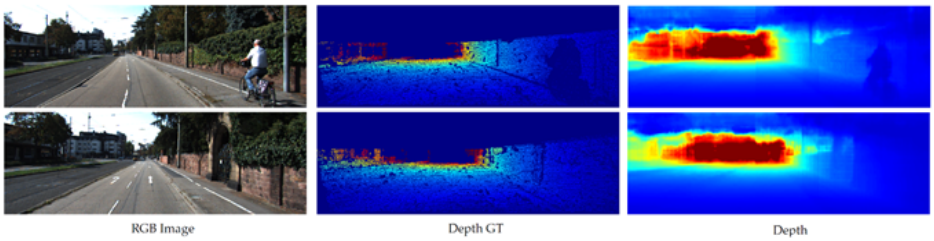


Fig. T1.2.8 The depth map is accurately generated, by using the novel stereo-based global optimization. The results also present a higher confidence, due to the geometrical constraints introduced by the method

Reference: VC. Miclea, S. Nedevschi, "Semi-Global Optimization for Classification-Based Monocular Depth Estimation", Proceedings of 2020 IEEE 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, and robust.

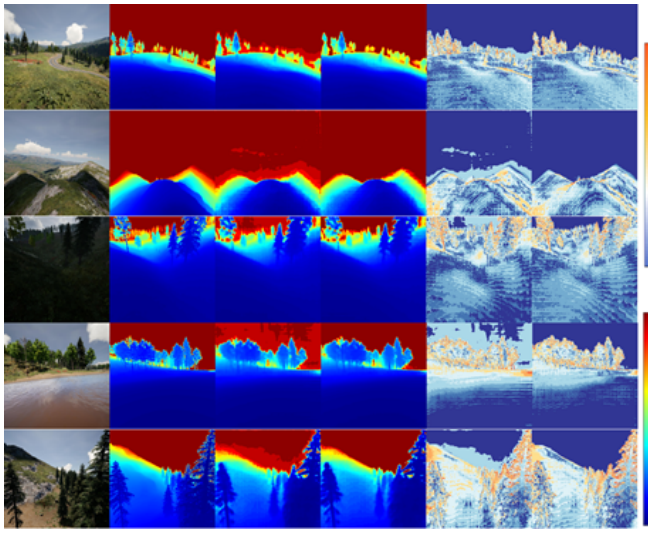


Fig. T1.2.9 MidAir images (synthetic) are used as input for the CNN, which produces highly accurate results (error images are presented on the last two columns)

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.

on

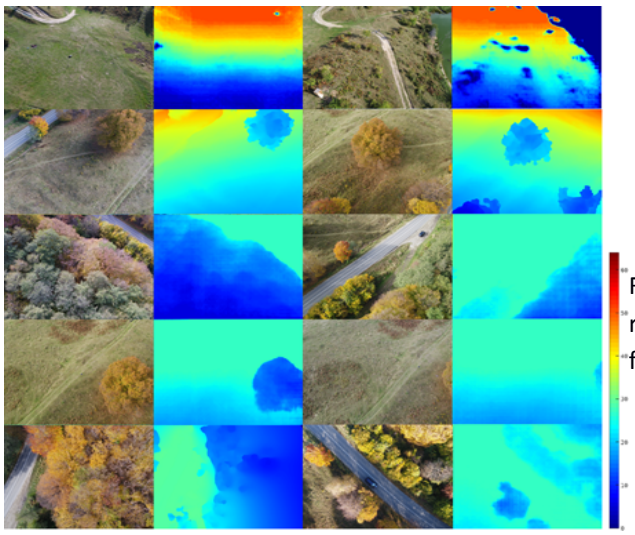


Fig. T1.2.10 The method produces very accurate depth results real-life images (Nadir and at various other angles), captured from a real drone, in various scenarios fields, forests)

***Reference**:* V. Miclea, S. Nedevschi, "Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception", in IEEE Transactions on Geoscience and Remote Sensing, Vol. 60, AN: 5602215, 2022

The article "WildUAV: Monocular UAV Dataset for Depth Estimation Tasks" presented the results of the training of several depth estimation systems based on deep learning using the WildUAV dataset. Four supervised algorithms were tested: a simple regression-based model, a classification-based model (depth being discretized into several classes), an ordinal regression model (discretized set of ordered values for depth), and a specially proposed model for aerial scenarios (combining the classical classification with the ordinal one), developed within this project and published in the article "Monocular Depth Estimation With Improved Long-Range Accuracy for UAV Environment Perception". Alternatively, three unsupervised algorithms were tested, as their flexibility, which stems from the lack of need for depth images during training, is an important advantage in some applications.
The performance of the supervised systems was compared between the training on synthetic aerial data and the training on real WildUAV data, the result favouring the second case. Moreover, the system specially developed for the air scenario was the most efficient, and benefited from the most important improvement after the use of the new data set. The qualitative evaluation shows that training on the mapping set together with the video one helps to improve the robustness of the system. The difficulty of estimating camera movement remains a limiting factor, especially in this aerial context where movements can occur over a much wider range of directions and magnitudes , requiring continued study.

***Reference**:* H. Florea, V.C. Miclea, S. Nedevschi, "WildUAV: Monocular UAV Dataset for Depth Estimation Tasks", in Proceedings of 17th *2021 IEEE International Conference Intelligent Computer Communication and Processing* (ICCP 2021).

In "Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation" a two-stage self-distillation framework trained on monocular video only, which does not require depth ground truth is introduced. In the first stage the teacher network is trained in a self-supervised regime and is further used to generate high-resolution scaled pseudo labels. In the second stage, the student network learns from scaled, consistent and improved pseudo labels for more accurate results. In unsupervised methods, there is the problem of the ambiguity of the depth scale, which is usually solved by the depth of ground truth. To solve this problem, we propose scaling the pseudo-depth by an automatic method. We also perform a pseudo-depth filter based on 3D consistency between consecutive frames, to eliminate errors.

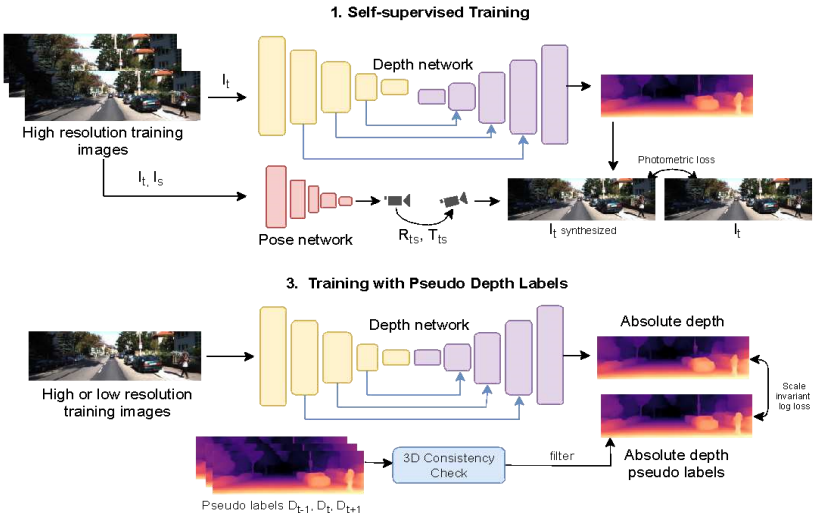


Fig. T1.2.11 Overview of the two-stage self-distillation framework for self-supervised depth estimation

Model	Resolution	AbsRel ↓	SqRel ↓	RMS ↓	RMSlog ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
Supervised	192 × 640	**0.095**	0.574	4.020	0.157	0.892	0.971	0.990
Self-supervised (gt scaling)	192 × 640	**0.104**	0.768	4.513	0.180	0.892	0.964	0.983
Self-supervised (automatic scaling)	192 × 640	0.108	0.795	4.655	0.192	0.878	0.959	0.981
Self-supervised (first stage - gt scaling)	320 × 1024	**0.101**	0.720	4.339	0.176	0.898	0.967	0.984
Self-supervised (first stage - automatic scaling)	320 × 1024	0.104	0.747	4.453	0.185	0.885	0.963	0.983
Pseudo-supervised (second stage)	192 × 640	0.100	0.661	4.264	0.172	0.896	0.967	0.985
Pseudo-supervised (second stage)	320 × 1024	**0.098**	0.721	4.415	0.180	0.890	0.965	0.984

Table 1.2.1 Comparison between the supervised, the unsupervised and the proposed pseudo-supervised method.

***Reference**:* A. Petrovai, S.Nedevschi, "Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation", *IEEE Computer Vision and Pattern Recognition* (CVPR 2022).

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.

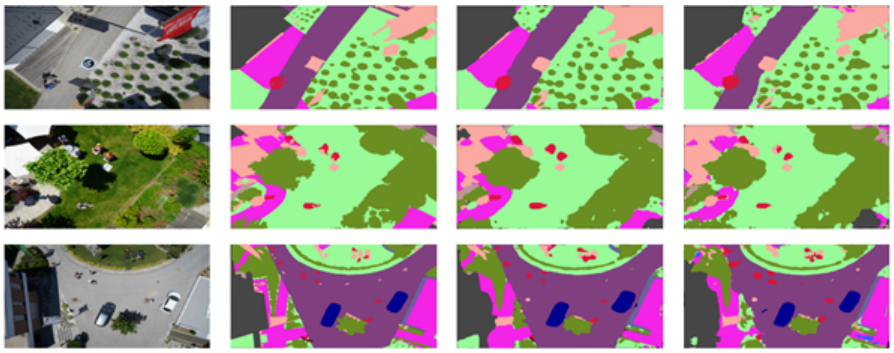


Fig. T2.1.1 The evaluation on the real drone dataset. From left to right: RGB image, ground truth for semantic annotation, inferred image when network is trained on the real dataset, inferred image when the network is trained on the merged dataset

***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.

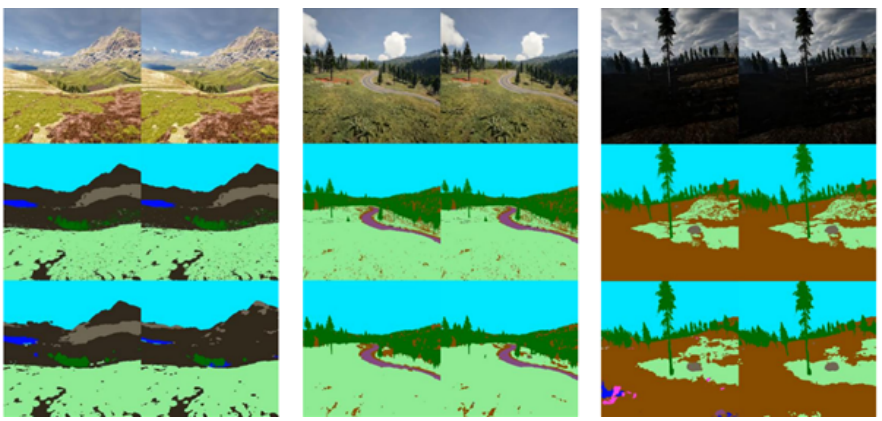


Fig. T2.1.2 Prediction results on the Mid-Air dataset, in 3 scenarios: mountain area, road in spring, and sunset in autumn. From top to bottom, the color image, the ground truth segmentation, and the semantic segmentation result. The first column is from MA50, while the second one – MA10.

***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.

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.

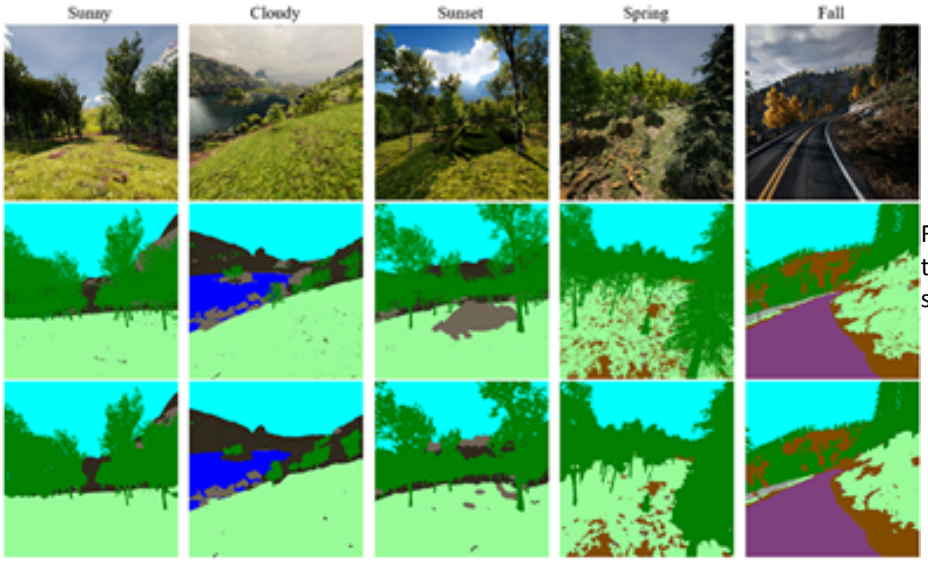


Fig. T2.1.3 Results of the framework on the test set, presented for the 5 different scenarios from Mid-Air.

***Reference**:* B. C. Z. Blaga and S. Nedevschi**,** ”Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences”, EUSIPCO 2021, Dublin, Ireland

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 „Multi-Task Network for Panoptic Segmentation in Automated Driving”. we introduced a panoptic head which is end-to-end trainable with the multi-task semantic and instance segmentation networks. 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.

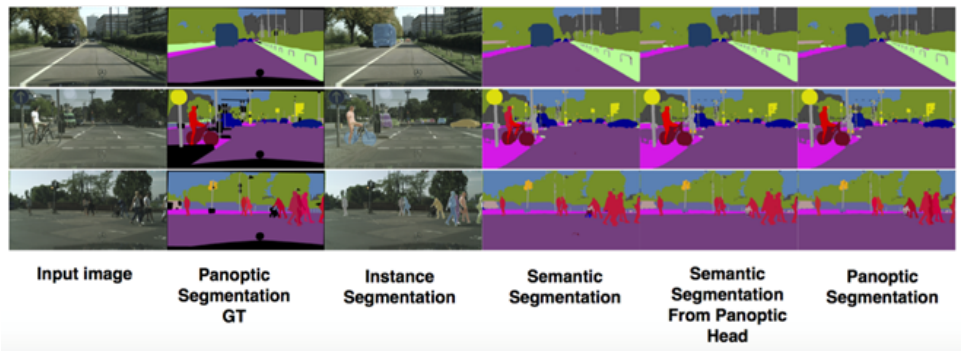


Fig. T2.2.1 End-to-end learning of, instance and semantic segmentation improves both semantic and instance segmentation results.

***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 "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.

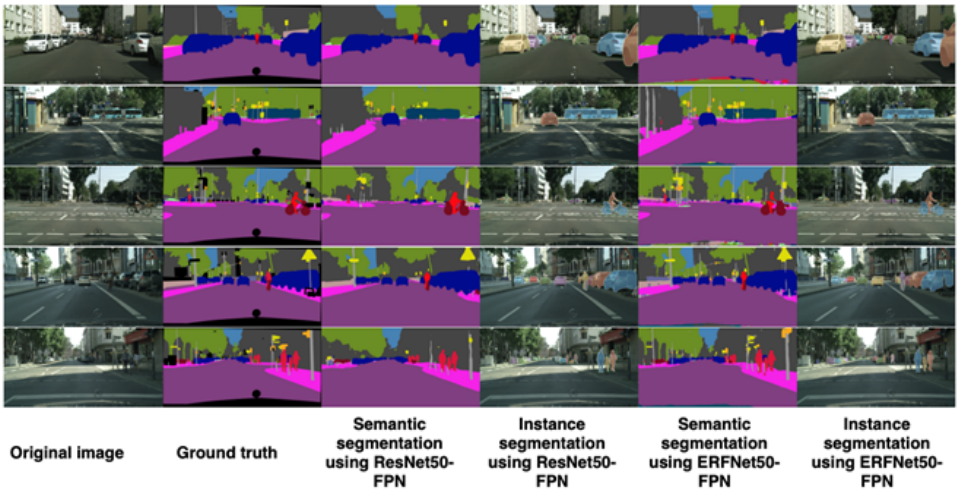


Fig. T2.2.2 The ResNet50-FPN is the baseline network and the ERFNet50-FPN is the proposed network. Compared to the baseline, the proposed solution increased the segmentation mIoU with 4.5%.

***Reference** :* A. Petrovai, S. Nedevschi, "Efficient instance and semantic segmentation for automated driving", *Proceedings 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.

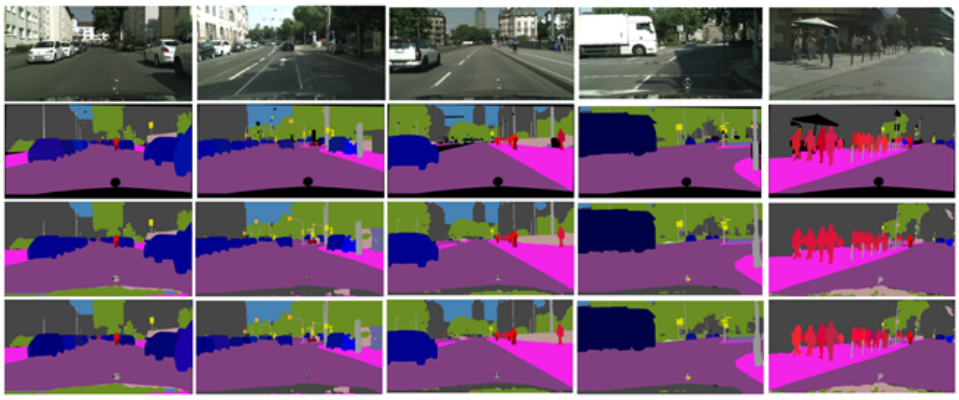


Fig. T2.2.3 From top to bottom: image, panoptic ground truth, semantic segmentation, panoptic segmentation. In the panoptic segmentation the color encodes the class and the instance identifier. Our network can correctly segment large and small scale objects and also occluded objects.

***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 "Fast Panoptic Segmentation with Soft Attention Embeddings". 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 at real-time performance in both inference speed and quality and achieves competitive results on the Cityscapes dataset.

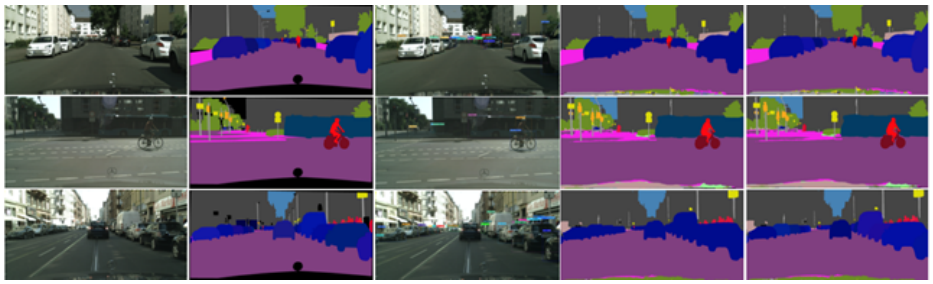


Fig. T2.2.4 From left to right: image, panoptic segmentation ground truth, object detection, semantic and panoptic segmentation. Our network can accurately segment object of various sizes and can handle difficult scenarios.

***Reference**:* A. Petrovai, S. Nedevschi, "Fast Panoptic Segmentation with Soft Attention Embeddings", *SENSORS*, Vol. 22, No. 3, AN: 783, FEB 2022.

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.

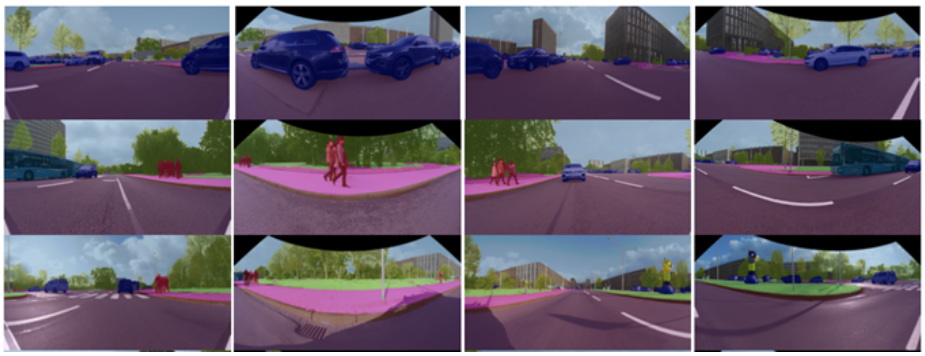


Fig. T2.2.5 Semantic segmentation of unwarped fisheye images. We process four images from the fisheye 160◦ horizontal field-of-view cameras which provide 360◦ coverage around the vehicle. Each camera views a different direction around the vehicle: front, right, rear and left.

***Reference**:* A. Petrovai, S. Nedevschi, "Semantic Cameras for 360-degree Environment Perception in Automated Urban Parking and Driving", in *IEEE Transactions on Intelligent Transportation Systems*, Early Access, March 2022.

Panoptic video segmentation simultaneously performs pixel-level semantic segmentation, instance-level segmentation, and instance tracking. We propose the VPS-Transformer network, with a hybrid architecture starting from the Panoptic-DeepLab network that performs panoptic image segmentation, which we expand with a new video module based on the Transformer architecture, which uses the attention mechanism to model spatio-temporal connections between images in a video. We propose methods for optimising the Transformer architecture, to obtain efficient processing of high resolution images.

Method	Backbone	PQ	VPQ	Time (ms)
VPSNet	ResNet50	62.7	56.1	770
ViP-DeepLab + MV	WR-41	67.9	59.9	> 400
Baseline - Panoptic DeepLab	ResNet50	63.0	52.0	86
VPS-Transformer (ours)	ResNet50	64.8	57.3	112
Baseline - Panoptic DeepLab	HRNet-W48	66.1	55.1	168
VPS-Transformer (ours)	HRNet-W48	67.6	59.8	185

Table 2.2.1. Comparison with the current state of research for panoptic segmentation of video sequences. We obtain a VPQ similar to Vip-DeepLab, but an inference time of 2x shorter.

***Reference**:* A. Petrovai, S. Nedevschi, "Time-Space Transformers for Video Panoptic Segmentation", *2022 IEEE Winter Conference on Applications of Computer Vision (WACV 2022)*, 4-8 January 2022, Waikoloa, Hawai, USA.

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 shows qualitative results .

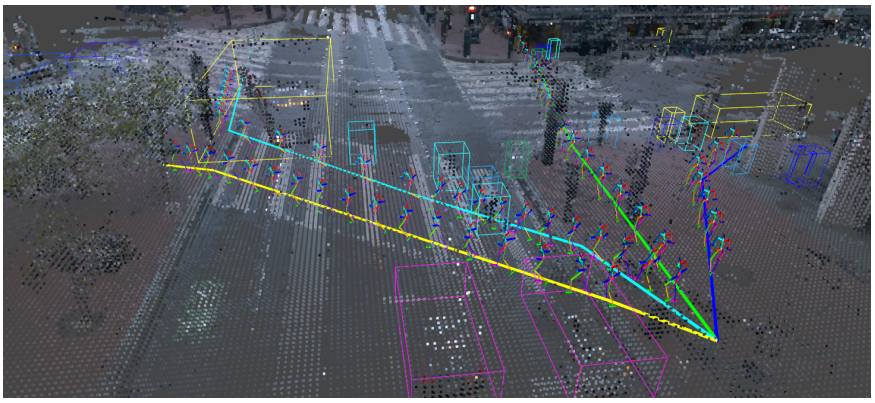


Fig. T3.1.1 SPL agent trajectories on the Waymo dataset, showing the pedestrian taking a number of different paths depending on how the agent history is initialized. Cars and other pedestrians are indicated with 3d bounding boxes. pedestrian trajectories. It should be noted that the collision-aware SPL 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.

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. To 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. Extensive experiments with the Panoptic multi-view setup, and for complex scenes containing multiple people, 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 below.

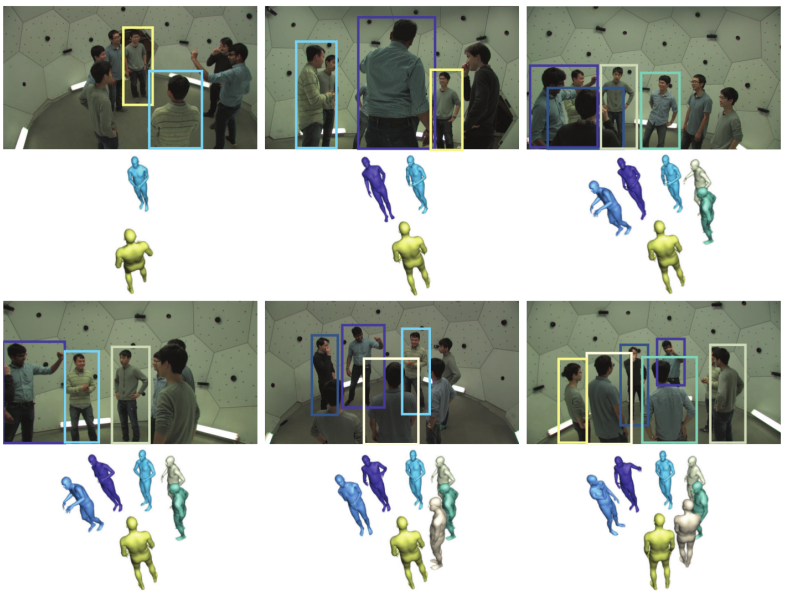


Fig. T3.1.2 Visualisation of how Pose-DRL performs multi-target pose estimation for an Ultimatum test scene. In this example the agent sees six viewpoints prior to automatically continuing to the next active-view. The mean error decreases from 358.9 to 114.6 mm/joint. Only two people are detected in 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.

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

An article published in Annual Conference on Robot Learning (CoRL2021, Proceedings of Machine Learning Research), Maria Priisalu, Aleksis Pirinen, Ciprian Paduraru si Cristian Sminchisescu discuss the inherent limitations the manually collected datasets impose on the variability of test cases and in particularly the difficulty to acquire challenging scenarios, e.g. ones involving collisions with pedestrians. A way to alleviate this is to consider automatic generation of safety-critical scenarios for autonomous vehicle (AV) testing. Existing approaches for scenario generation use heuristic pedestrian behavior models. We instead propose a framework that can use state-of-the-art pedestrian motion models, which is achieved by reformulating the problem as learning where to place pedestrians such that the induced scenarios are collision prone for a given AV. Our pedestrian initial location model can be used in conjunction with any goal driven pedestrian model which makes it possible to challenge an AV with a wide range of pedestrian behaviors – this ensures that the AV can avoid collisions with any pedestrian it encounters. We show that it is possible to learn a collision seeking scenario generation model when both the pedestrian and AV are collision avoiding. The initial location model is conditioned on scene semantics and occlusions to ensure semantic and visual plausibility, which increases the realism of generated scenarios. Our model can be used to test any AV model given sufficient constraints. The next figure shows simultaneously trained trajectories, using an initialization supposed to lead to collision, as well as one that doesn't lead to collision because the self-driving car accelerates:

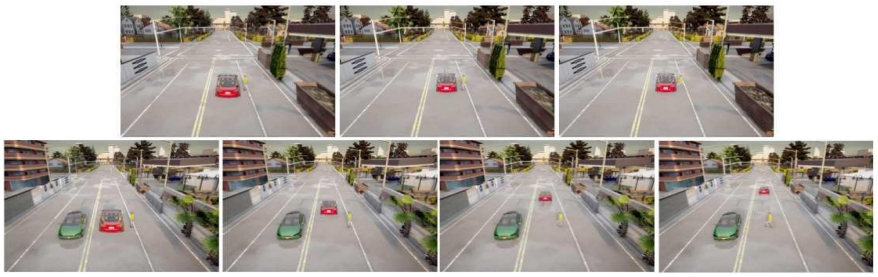


Fig. T3.1.3 Sample trajectories of the Simultaneous-μ, ρ model, sub-sampled at 5 frames from frame 0. First row: The AV changes speed thus causing the pedestrian to incorrectly estimate the AV's motion and walk into the AV. Second row: The pedestrian waits for the AV to pass before crossing the road.

Reference: Maria Priisalu, Aleksis Pirinen, Ciprian Paduraru, Cristian Sminchisescu. Generating Scenarios with Diverse Pedestrian Behaviours for Autonomous Vehicle Testing, In Proceedings of Machine Learning Research (Proc. of CoRL 2021)

Task 3.2. Representations and Methods for Efficient Computation

We studied 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 developed 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 evaluated 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.

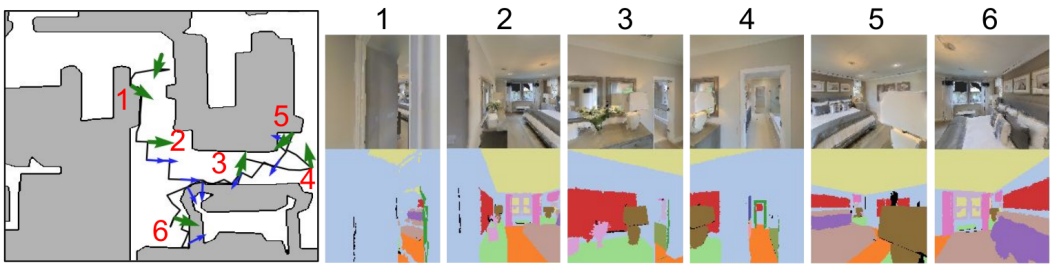


Fig T3.2.1 The first six requested annotations by the RL-agent in a room from the test set. Left: Map showing the agent's trajectory and the six first requested annotations (green arrows). The initially given annotation is not indicated with a number. Blue arrows indicate Collect actions. Right: For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

***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 Speople 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, the agent 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 the agent 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. T3.2.2 below.

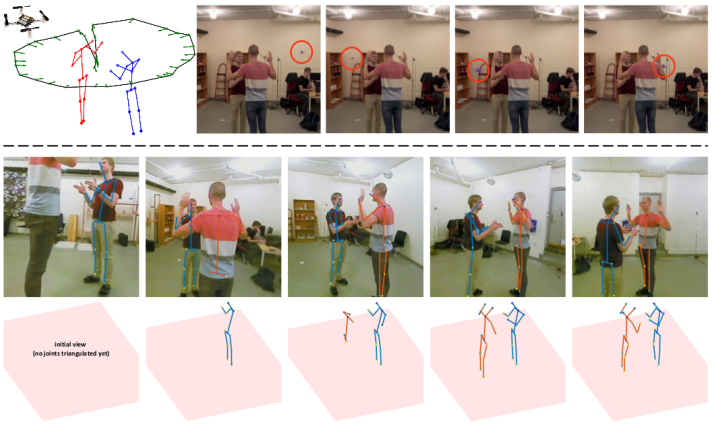


Fig. T3.2.2 Proof-of-concept experiment illustrating how the agent can be connected to an active drone observer to reconstruct 3d poses from informative viewpoints. Above the dashed line to the left we show the drone's loop (the sharp peak is due to take-off and landing), with sampled camera locations as green arrows. 3d pose reconstructions obtained by triangulating from all 33 sampled camera locations are also shown. The drone used is shown in the very top left corner; it can be used safely due to its small size and weight. Sample drone locations are also shown above the line (highlighted with red circles in images). Below the line we show views seen by the agent 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. 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 based on traditional structure from motion and multi-view-stereo techniques, 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.

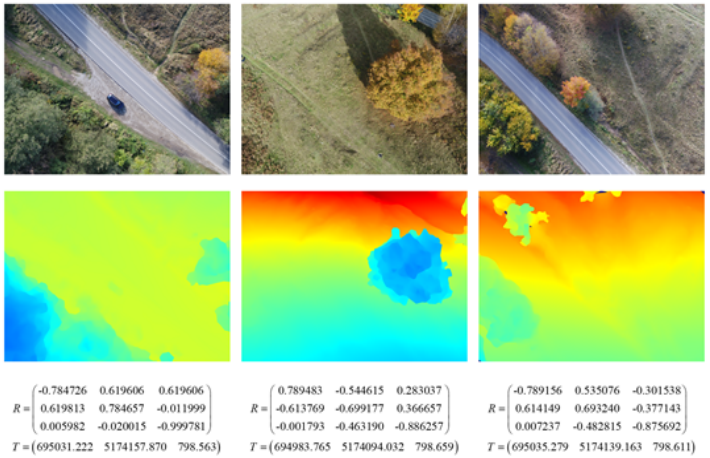


Fig. T4.1.1. Examples of the dense depth maps (second row) obtained by reprojecting the 3D mesh onto the original images (first row) for which the precise pose is available (third row)

***Reference**:* H. Florea, V.C. Miclea, S. Nedevschi, "WildUAV: Monocular UAV Dataset for Depth Estimation Tasks" in ICCP '21

Ground-following trajectory planner. A trajectory planner was developed for plotting new UAV flight trajectories based on reconstructed 3D maps of an area. Ground surface elevation information is exploited in order to generate trajectories which maintain a constant height above the terrain, unlike standard constant-altitude trajectories. This is especially useful when recording data above areas with varying terrain profiles at low altitudes, as it allows capturing images where scene objects maintain relative constant appearance.

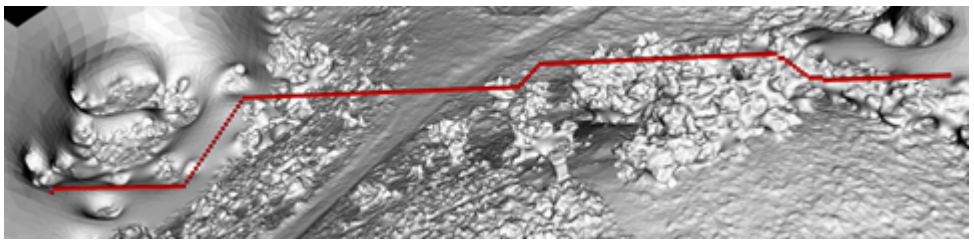


Fig. T4.1.2 : The path resulting from the A* algorithm

***Reference**:* V. Chiciudean F. Oniga, "Pathfinding in a 3D Grid for UAV Navigation", submitted to IEEE ICCP 2022

Task 4.2: Integration of visual navigation and scene understanding.

Semantics and geometry fusion. 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 optimised 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.

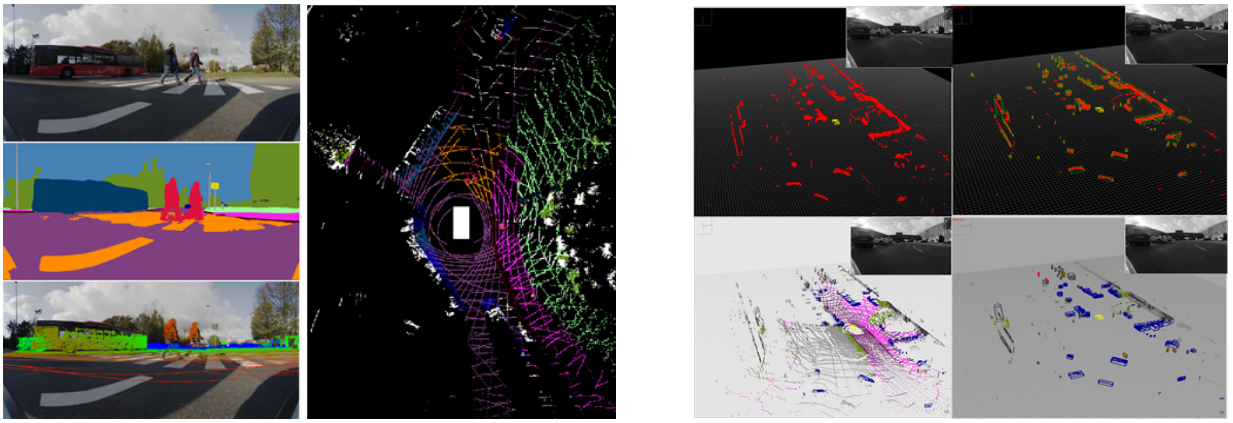


Fig. T4.2.1 Left: Low level fusion concept and Spatio-Temporal and Appearance Based Representation (STAR), Right: Enhanced perception based on the Spatio-Temporal and Appearance Based Representation (STAR)

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

Integration of panoptic segmentation with 3D geometry and objects tracking in aerial and automotive environments. Depth-aware video panoptic segmentation tackles the inverse projection problem of restoring panoptic 3D point clouds from

video sequences, where the 3D points are augmented with semantic classes and temporally consistent instance identifiers. We propose a novel solution with a multi-task network that performs monocular depth estimation and video panoptic segmentation. Since acquiring ground truth labels for both depth and image segmentation has a relatively large cost, we leverage the power of unlabeled video sequences with self-supervised monocular depth estimation and semi-supervised learning from pseudo-labels for video panoptic segmentation. To further improve the depth prediction, we introduce panoptic-guided depth losses and a novel panoptic masking scheme for moving objects to avoid corrupting the training signal. Extensive experiments on the Cityscapes-DVPS and SemKITTI-DVPS datasets demonstrate that our model with the proposed improvements achieves competitive results and fast inference speed. The model proved its usability even on UAVid aerial data set providing simultaneously instance level segmentation of the aerial images, tracking of the moving objects and a 3D reconstruction and semantic segmentation of the environment representing 3D point cloud.

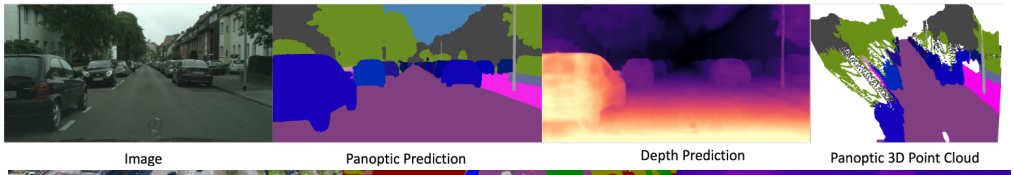


Fig. T4.2.2 Depth-aware video panoptic segmentation

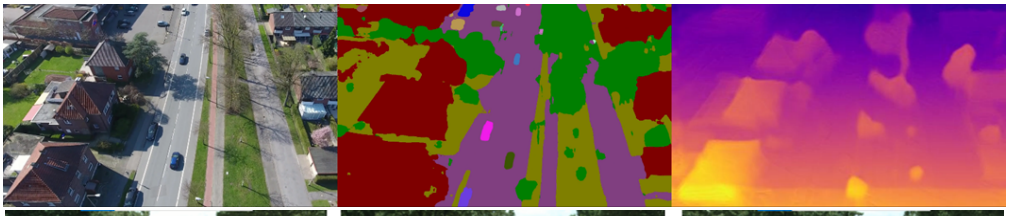


Fig. T4.2.3 Depth-aware video panoptic segmentation from aerial images

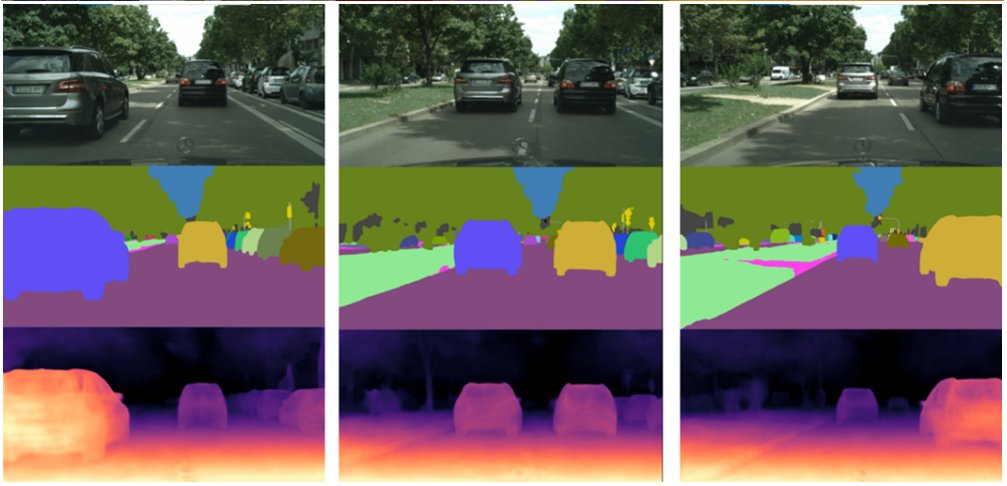
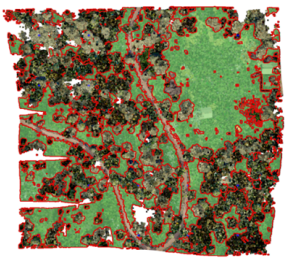


Fig. T4.2.4 Depth-aware video panoptic segmentation in automotive environment

Reference: A. Petrovai, S.Nedevschi, “MonoVPS: A Self-Supervised Monocular Depth Estimation Approach to Depth-Aware Video Panoptic Segmentation”, under evaluation ECCV 2022

Estimation of deforested areas. Starting from the virtual environment in the simulator, we developed a methodology to describe the degree of afforestation of the area of interest. The first step is to map the forest, which is done by recording color and semantic images from a drone flying over the area. The second step is to build a cloud of 3D dots with color information and semantic segmentation, for which we used the Open3D library. By temporarily merging the records, a map is obtained based on which we can extract the percentage of healthy trees, but we also have information about the degree of deforested trees and deforested area. The last step is to apply a morphological expansion operation through which we obtained a denser map of semantic information, based on which we extracted the contours of deforested areas, results that can be seen below.



Deforestation: 53,70%
Healthy trees: 39.73%
Fallen trees: 0.48%

Fig. T4.2.5 Extraction of treeless regions and information on the degree of deforestation relative to the analysed area.

Reference: Z. Blaga, S. Nedevschi, "Forest Inspection Dataset for Aerial Semantic Segmentation and Depth Estimation", submitted to TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING ON JUNE 2022

Autonomous navigation based on obstacle avoidance in forest environments. We develop an autonomous navigation application integrated into the AirSim simulation environment which issues UAV control commands based on information from the forward-looking camera. Depth estimation is carried out by a Self-Supervised Monocular Depth Estimation CNN that outputs up-to-scale results for each frame, which can then be converted to metric results using a median scaling factor obtained initially during ground-truth evaluation (a dynamic scaling factor adjustment was implemented for improving the results). The navigation follows a trajectory described as 3D waypoints, with height and yaw corrections applied based on ground and obstacle distances. Obstacle avoidance is carried out by monitoring a region of interest on the depth image and issuing course corrections when the distance to the object directly ahead is below a safety threshold. The correction angle is computed based on the most suitable vertical image strip (i.e. the one showing the highest possible distance in the depth image).

3. Impact & significant results

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

Task 1.1: Deep Learning of Graph Matching under Global Constraints

A best paper honourable mention awarded paper at CVPR 2018, "**Deep Learning of Graph Matching**", a top A* conference, presents significant methodological results. Specifically, 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

"**Deep Network for the Integrated 3D Sensing of Multiple People in Natural Images**" (Neurips 2018, A* conference) proposes a novel, feedforward deep network, supporting different supervision regimes, that predicts the 3d pose and shape of multiple people in monocular images. The main difficulty addressed refers to the formulation and integration of localization and grouping people into the network, as a binary linear integer program, to be solved 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.

"**Human Synthesis and Scene Compositing**" (AAAI 2020, A* conference) and presents a framework that realistically synthesises a photograph of a person, in any given pose and shape, and blends it veridically with a new scene, while obeying 3d geometry and appearance statistics. The significant results 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 may have an important impact in photo editing, fashion virtual try-on, or realistic data augmentation used for training large scale 3d human sensing models.

One of the significant results of "**Three-dimensional Reconstruction of Human Interactions**" (CVPR 2020, A* conference), 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. A second significant result, with an important potential impact, is represented by the datasets collected to prove the value of our proposed methodology (*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). Models and data are made available for research at http://vision.imar.ro/ci3d.

To overcome some of the shortcomings of existing, self-contact agnostic, 3d reconstruction methods, "**Learning Complex 3D Human Self-Contact**" (AAAI 2021, A* conference), presents 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. The paper also presents two large scale datasets collected to train models and for large-scale quantitative evaluation: *HumanSC3D*, 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, and *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. Other significant results include: (a) a quantitative and qualitative proof of metrically more accurate and perceptually veridical 3d reconstructions based on self-contact signatures, and (b) a foundation for a large class of applications that would benefit from accurate 3d self-contact representations (eg., infection health monitoring when hands touch parts of the face (mouth, nose, eyes) in hospitals or during a pandemic, or subtle behavioural understanding of gestures for robot-assisted therapy of children with autism).

"REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People under Weak Supervision" (NeurIPS 2021, A* conference) addresses the limitations of the current 3D human reconstruction methods that either treat each person independently, ignoring most of the context, or reconstruct people jointly, but cannot recover interactions correctly when people are in close proximity. In this work, we introduce REMIPS, a model for 3D Reconstruction of Multiple Interacting People under Weak Supervision that can reconstruct a variable number of people directly from monocular images. At the core of our methodology stands a novel transformer network that combines unordered person tokens (one for each detected human) with positional-encoded tokens from image features patches into a novel unified model for self- and interpenetration-collisions based on a mesh approximation computed by applying decimation operators. Our main contributions can be summarized as follows: (1) fast, accurate and unified 3d self collision and interpenetration models for multiple people; (2) a novel vision transformer architecture to predict 3D pose and shape for multiple people; (3) weakly supervised models which do not require 3D annotations during training; (4) state of the art results on challenging datasets, with favorable performance compared to competing predictive or optimization-based methods.

To address the lack of robustness of supervised monocular depth estimation (MDE) methods when compared to stereo or structure from motion than use geometric constraints to derive the depth, we proposed in the article "**SGM-MDE: Semi-global optimization for classification-based monocular depth estimation**" (**International Conference on Intelligent Robots and Systems, IEEE IROS, 2020), 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. The semi-global stereo optimization 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.

"**A unified method for improving long-range accuracy of stereo and monocular depth estimation algorithms**" (**2020 IEEE Intelligent Vehicles Symposium),** introduces 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 neighbouring feature vectors and then 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 works for both stereo and monocular depth perception algorithms that output a depth in a discrete setting (most suitable for real-time applications). 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.

We further improved supervised MDE in the article "**Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception**" (**IEEE Transactions on Geoscience and Remote Sensing 2021),** where we introduced a novel approach capable to work on complex aerial images, captured from a medium distance, in a variety of scenarios. **The novel CNN proposed contains a Darknet-based feature extractor, that is both lightweight and properly built to capture aerial information. 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 also benefits from the aforementioned refinement proposals, increasing robustness by using the global optimization and dealing with objects at large distances. The proposed CNN provides the most accurate results in the aerial image category (an average error reduction of at least 2 meters over previous work performed on Kitti benchmark). 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.

"**Real-time Semantic Segmentation-based Stereo Reconstruction**" (**IEEE Transactions on Intelligent Transportation Systems)** proposes a novel real-time stereo reconstruction solution accounting for both geometry and semantics to produce an accurate depth map by taking advantage of both traditional stereo methods and pure learning-based solutions, while addressing their respective weaknesses. To this end, we introduced **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**. A semantic map of the scene is generated by using a CNN and subsequent stereo pipeline steps (cost computation, aggregation, optimization and refinement) are tailored to incorporate scene information from the semantic map to enhance the results. A novel real-time CNN is proposed for the refinement step. The CNN simultaneously extracts features from the RGB image, the uncorrected depth map and semantic segmentation, and cleverly combines them towards an improved depth map, using regression. The new method produces the best real-time stereo reconstruction results on Kitti 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.

"**Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation**" (CVPR 2022) proposes a two-stage self-distillation framework trained on monocular video only, which does not require depth ground truth. In the first stage the teacher network is trained in a self-supervised regime and is further used to generate high-resolution scaled pseudo labels. In the second stage, the student network learns from scaled, consistent and improved pseudo labels for more accurate results. In unsupervised methods. To solve the problem of the ambiguity of depth scale, we proposed scaling the pseudo-depth by an automatic method. We checked the generalisation capability of our model trained on KITTI, without any fine-tuning on Cityscape. Compared to Monodepth2 and other competing methods, we achieve better scores across all metrics.

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

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

A novel image panoptic segmentation algorithm to tackle the task of semantic perception of the environment is proposed in "**Multi-task Network for Panoptic Segmentation in Automated Driving**" (2019 IEEE Intelligent Transportation Systems Conference) as **an improved segmentation head on top of the Feature Pyramid Network of Mask R-CNN**. **Our second**

contribution is 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. The model evaluation on the Cityscapes dataset for semantic segmentation showed 73.3 mIoU from the segmentation head, with a further improvement of the panoptic head to 75.4 mIoU.

Mask R-CNN based approaches for panoptic segmentation are accurate but not suitable for real-time processing. "**Real-Time Panoptic Segmentation with Prototype Masks for Automated Driving**" (2020 IEEE Intelligent Vehicles Symposium) proposes a fast fully CNN 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. We extend the single-shot and anchor-free FCOS object detector with a lightweight semantic segmentation decoder, that is used for stuff class mask prediction**. **A novel panoptic segmentation network design 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. Evaluation on the challenging Cityscapes dataset achieves state-of-the-art results at 76.9% mIoU and 57.3 PQ, outperforming previous solutions both in terms of accuracy and speed. Moreover, the fully convolutional nature of the network facilitates deployment for robotic applications.

In the paper "**SAPSNet: A Soft Attention Panoptic Segmentation Network**", 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**. The evaluation of our solution outperforms on the Cityscapes dataset shows on par or better results compared to previous work.

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

Task 3.1. Direct and Inverse Optimal control

"**Semantic Synthesis of Pedestrian Locomotion**" (ACCV 2020), proposes 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. The significant results described in the paper are as follows: (a) an articulated 3d pedestrian motion generator that conditions its predictions on both the scene semantics and human locomotion dynamics; (b) 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) extensive evaluations on Cityscapes, Waymo and CARLA that show 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.

"**Deep Reinforcement Learning for Active Human Pose Estimation**" (AAAI 2021, A* conference) presents a fully trainable deep reinforcement-learning based active vision model for human pose estimation. The significant results are the accuracy of the estimates and the capacity to select an adaptively selected number of informative views which results 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. The paper presents fundamental methodology required for future applied research.

Task 3.2. Representations and Methods for Efficient Computation

"**Embodied Visual Active Learning for Semantic Segmentation**" (AAAI 2021, A* conference) presents methodology for visual active learning for semantic segmentation where an agent is set to explore a 3d environment aiming to acquire visual scene understanding by actively selecting views for which to request annotation. The most significant results of the work are: (a) a study of 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 neighbourhood 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) a variety of methods, both learnt and prespecified, to tackle our task in the context of semantic segmentation; (d) extensive evaluation in a photorealistic 3d environment that shows that a fully learnt method outperforms comparable pre-specified ones.

"**Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstruction**" (NIPS 2019, A* conference), presents a deep reinforcement learning-based agent to actively reconstruct 3d poses from 2d estimates via triangulation. The most significant results 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 the agent to a physical drone observer.

A4. Systems Optimization and Integration

Task 4.1: Simultaneous localization and semantic mapping

"WildUAV: Monocular UAV Dataset for Depth Estimation Tasks" presents an aerial imaging dataset created using a DJI Matrice 210 V2 RTK drone with over 1800 high-resolution images alongside video sequences from multiple flights over forest and open terrains, for which accurate positioning data is available. The dataset features textured 3D meshes 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.

Ground-following trajectory planner. **"Pathfinding in a 3D Grid for UAV Navigation"** presents a novel trajectory planner for plotting new UAV flight trajectories based on reconstructed 3D maps of an area. Ground surface elevation information is exploited to generate trajectories which maintain a constant height above the terrain, unlike standard constant-altitude trajectories. This is especially useful when recording data above areas with varying terrain profiles at low altitudes.

Task 4.2: Integration of visual navigation and scene understanding

Semantics and geometry fusion. "**Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion**" presents 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 optimised 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.

Integration of panoptic segmentation with 3D geometry and objects tracking in aerial and automotive environments. **"MonoVPS: A Self-Supervised Monocular Depth Estimation Approach to Depth-Aware Video Panoptic Segmentation"** proposes a novel solution for depth-aware video panoptic segmentation with a multi-task network that performs monocular depth estimation and video panoptic segmentation, by leveraging the power of unlabeled video sequences with self-supervised monocular depth estimation and semi-supervised learning from pseudo-labels for video panoptic segmentation. To further improve the depth prediction, we introduce panoptic-guided depth losses and a novel panoptic masking scheme for moving objects to avoid corrupting the training signal. The model proved its usability even on UAVid

aerial data set providing simultaneously instance level segmentation of the aerial images, tracking of the moving objects and a 3D reconstruction and semantic segmentation of the environment representing 3D point cloud.

***Estimation of deforested areas. "*Forest Inspection Dataset for Aerial Semantic Segmentation and Depth Estimation"** develops a methodology to describe the degree of deforestation of an area of interest, based on a simulated virtual environment. The methodology maps the forest (by using color and semantic images collected by a drone) and builds a cloud of 3D dots with color information and semantic segmentation. By temporarily merging the records, a map is obtained based on which the percentage of healthy trees can be found, as well as the information about the degree of deforested trees and deforested area. A morphological expansion operation helps obtain a denser map of semantic information used to extract the contours of deforested areas.

Autonomous navigation based on obstacle avoidance in forest environments. We developed an autonomous navigation application integrated into the AirSim simulation environment which issues UAV control commands based on information from the forward-looking camera. Depth estimation is carried out by a Self-Supervised Monocular Depth Estimation CNN that yields up-to-scale results for each frame, which can then be converted to metric results using a median scaling factor obtained initially during ground-truth evaluation (a dynamic scaling factor adjustment was implemented to improve the results). The navigation follows a trajectory described as 3D waypoints, with height and yaw corrections applied based on ground and obstacle distances. Obstacle avoidance is carried out by monitoring a region of interest on the depth image and issuing course corrections when the distance to the object directly ahead is below a safety threshold. The correction angle is computed based on the most suitable vertical image strip (i.e. the one showing the highest possible distance in the depth image).

4. Result indicators

4.1 Articles: 10 (3 under evaluation/7 published)

[1] H. Petzka and C. Sminchisescu, Non-attracting Regions of Local Minima in Deep and Wide Neural Networks, Journal of Machine Learning Research, July 2021. (Impact factor 2018: 4.091)
[2] Vlad Miclea, Sergiu Nedevschi, Real-Time Semantic Segmentation-Based Stereo Reconstruction, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, April 2020 (Impact factor 6.319)
[3] M.P. Muresan, I. Giosan, S. Nedevschi, Stabilisation and Validation of 3D Object Position Using Multimodal Sensor Fusion and Semantic Segmentation, SENSORS, vol. 20, no. 4, article number 1110, FEB 2020, (Impact factor 3.57)
[4] Vlad Miclea, Sergiu Nedevschi. Monocular Depth Estimation with Improved Long-range Accuracy for UAV Environment Perception, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, March 2021(Impact factor 2020: 5.855)
[5] A. Petrovai, Sergiu Nedevschi. Fast Panoptic Segmentation with Soft Attention Embeddings. Sensors, Jan. 2022 (IF: 3.57)
[6] H. Florea, A. Petrovai, I. Giosan, F. Oniga, R. Varga, S. Nedevschi, "Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion", under evaluation IEEE Transactions on Intelligent Transportation Systems, (IF: 6.319)
[7] Andra Petrovai, Sergiu Nedevschi . Semantic Cameras for 360-degree Environment Perception in Automated Urban Driving. IEEE Transactions on Intelligent Transportation Systems (Early Access), March 2022 (Impact factor 6.319)
[8] M.P. Muresan, S. Nedevschi, R. Danescu, "Robust Data Association using Fusion of Data-Driven and Engineered Features for Real Time Pedestrian Tracking in Thermal Images", SENSORS, Vol. 21 Issue 23, AN 8005, NOV 2021, (Impact factor 3.57)
[9] VC Miclea, S. Nedevschi, Dynamic Semantically Guided Monocular Depth Estimation for UAV Environment Perception, submitted to TO TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING ON JUNE 2022, (Impact factor 2020: 5.855)
[10] Z. Blaga, S. Nedevschi, Forest Inspection Dataset for Aerial Semantic Segmentation and Depth Estimation, submitted to TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING ON JUNE 2022 (Impact factor 2020: 5.855)

4.2 Patents: 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

4.3 Conferences: 33 (3 under evaluation / 30 published)

Selected conferences (for a complete list of publications, please see http://vision.imar.ro/sepca/publications.html):

[1] 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
[2] 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 32nd Conference on Neural Information Processing Systems, NIPS 2018 - A* conference
[3] 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
[4] 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
[5] 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
[6] 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
[7] 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 - A* conference
[8] 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
[9] 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
[10] M. Priisalu, A. Pirinen, Ciprian Paduraru, Cristian Sminchisescu. Generating Scenarios with Diverse Pedestrian Behaviours for Autonomous Vehicle Testing, In Proceedings of Machine Learning Research (Proceedings of CoRL 2021) - A conference
[11] Mihai Fieraru, Teodor Alexandru Szente, Eduard Gabriel Bazavan, Vlad Olaru and Cristian Sminchisescu. REMIPS: Physically Consistent 3D Reconstruction of Multiple Interacting People under Weak Supervision, in Advances in Neural Information Processing Systems 34 (NeurIPS 2021) - A* conference
[12] 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.
[13] 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
[14] 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[15] 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
[16] 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
[17] 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
[18] 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), Oct. 19–Nov. 13, 2020, Las Vegas, SUA, - B conference
[19] 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
[20] B.C.Z. Blaga, S. Nedevschi, “Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences”, EUSIPCO 2021, Dublin, Ireland
[21] A. Petrovai, S. Nedevschi, “Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation ”, 2022 Conference on Computer Vision and Pattern Recognition (CVPR 2022). 19-24 June 2022, New Orleans, SUA - A* conference
[22] A. Petrovai, S. Nedevschi, “Time-Space Transformers for Video Panoptic Segmentation”, 2022 IEEE Winter Conference on Applications of Computer Vision (WACV 2022), 4-8 January 2022, Waikoloa, Hawai, USA - A conference
[23] A. Petrovai, S.Nedevschi, “MonoVPS: A Self-Supervised Monocular Depth Estimation Approach to Depth-Aware Video Panoptic Segmentation”, under evaluation ECCV 2022 - A* conference

Date 30.06.2022 **Project manager** *Cristian Sminchisescu*