Scientific report

on project implementation status between January-December 2019

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

Financing: UEFISCDI, Nr. 9/2018

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

Integrated Semantic Visual Perception and Control for Autonomous Systems

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

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

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

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

A1. Deep 3D Reconstruction

Task 1.2: Deep Structured Geometric Models with Semantics

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

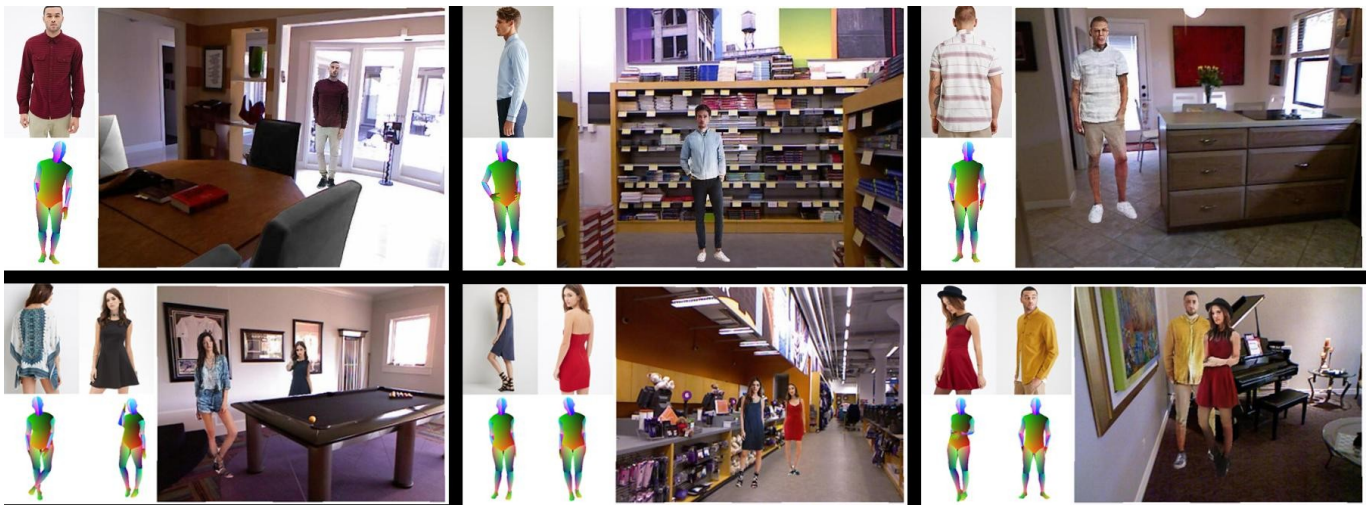

Fig T1.2.1 Sample images generated by our framework. For each example, we show the source image, the target 3d body mesh, and a scene with a geometrically plausible placement of the synthesized person. Please note that our framework allows for a positioning behind various objects, and the insertion of multiple people without breaking any geometrical scene properties.

Reference

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

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

Fig. T1.2.2 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”, submitted to the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

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.

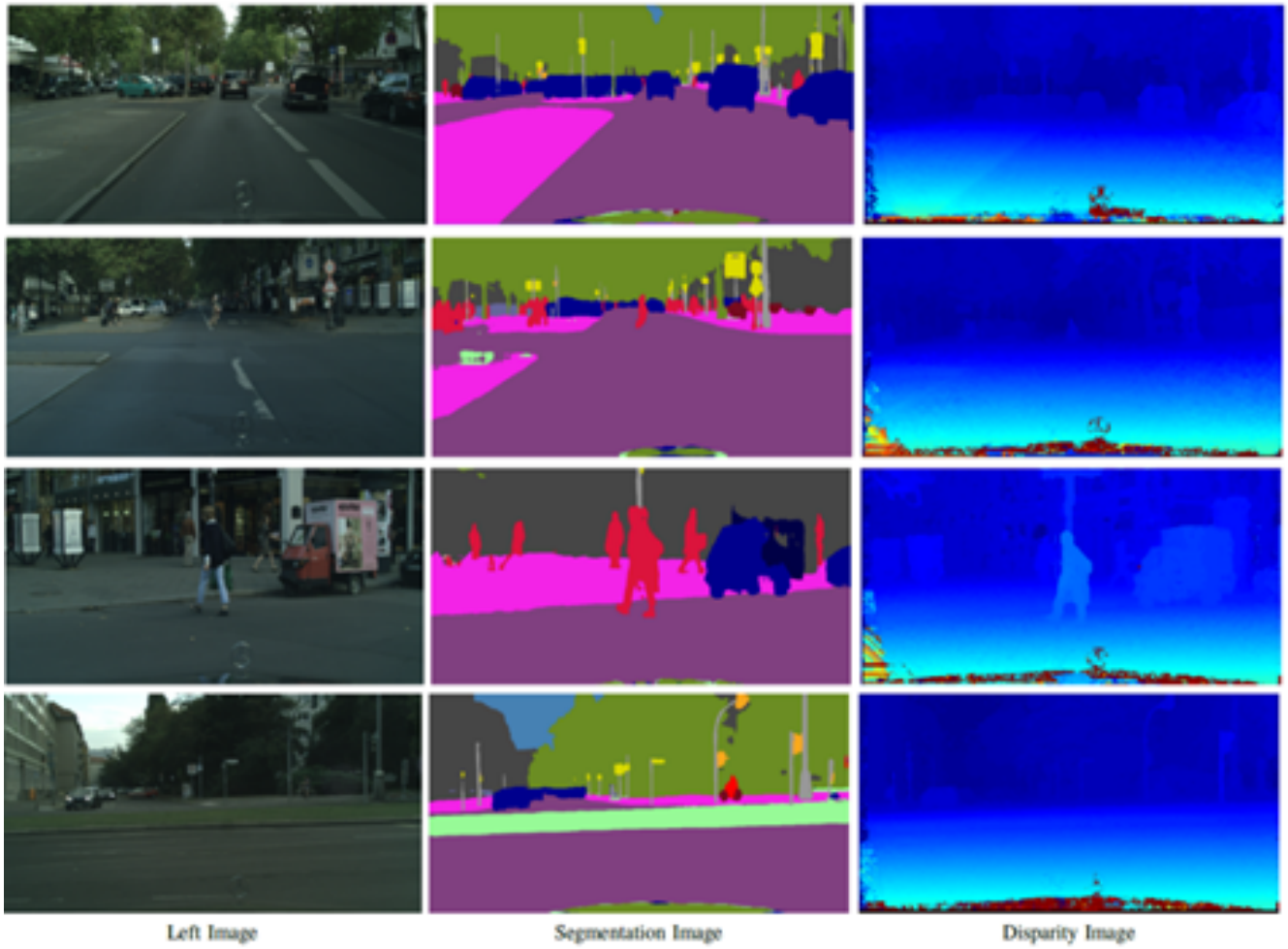

Fig. T1.2.3 Results from the paper “Real-Time Semantic Segmentation-Based Stereo Reconstruction”. The disparity image (from Cityscapes dataset) is obtained by enhancing the stereo pipeline with information from the semantic segmentation. The proper shape of the objects is clearly captured in all scenarios

Reference

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

A2. Visual Recognition and Localization

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

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

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

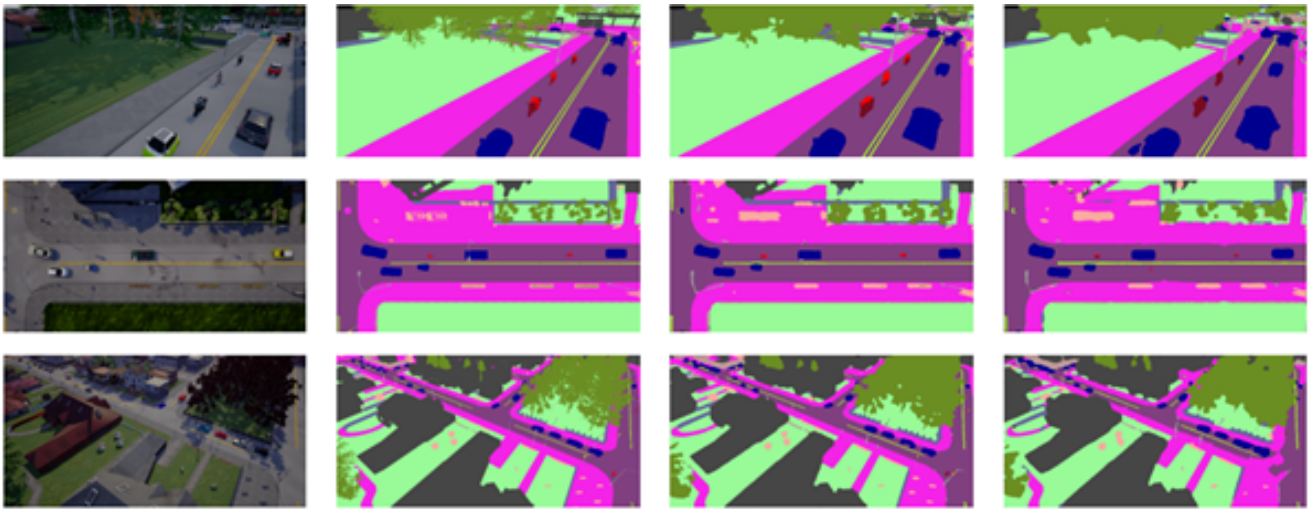

Fig. T2.1.1 Results from the paper **“**Semantic Segmentation Learning for Autonomous UAVs using Simulators and Real Data**”.** The evaluation of the semantic segmentation on the synthetic dataset. From left to right: RGB image, ground truth for semantic annotation, inferred image when network is trained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

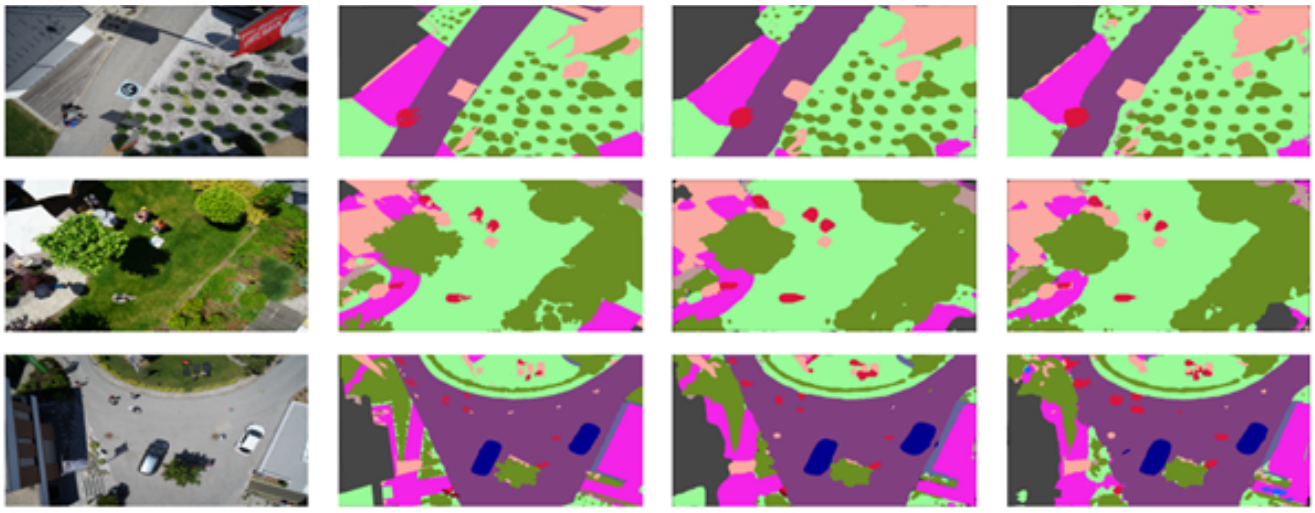

Fig. T2.1.2 Results from the paper **“**Semantic Segmentation Learning for Autonomous UAVs using Simulators and Real Data**”.** The evaluation of the semantic segmentation on the real drone dataset. From left to right: RGB image, ground truth for semantic annotation, inferred image when network is trained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

Reference

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

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

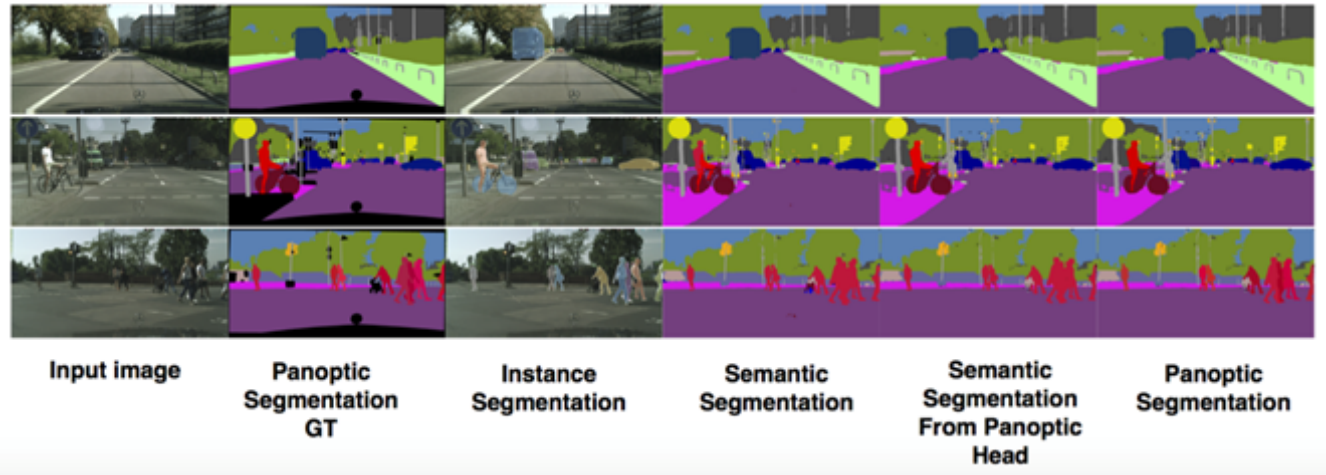

Fig. T2.2.1 Results from the paper „Multi-Task Network for Panoptic Segmentation in Automated Driving”. End-to-end learning of, instance and semantic segmentation improves both semantic and instance segmentation results.

Reference

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

Real-time performance is crucial for many robotic applications that perform online environment perception. Although these two-stage semantic and instance segmentation methods are accurate, they are not suitable for real-time processing. In the paper

“Efficient instance and semantic segmentation for automated driving”we study how to speed up two-stage semantic and instance networks and propose a fast and efficient two-stage network that can reach better accuracy than the slower baseline. Our proposed network features a backbone with factorized convolutions and dilated convolutions for increased accuracy.

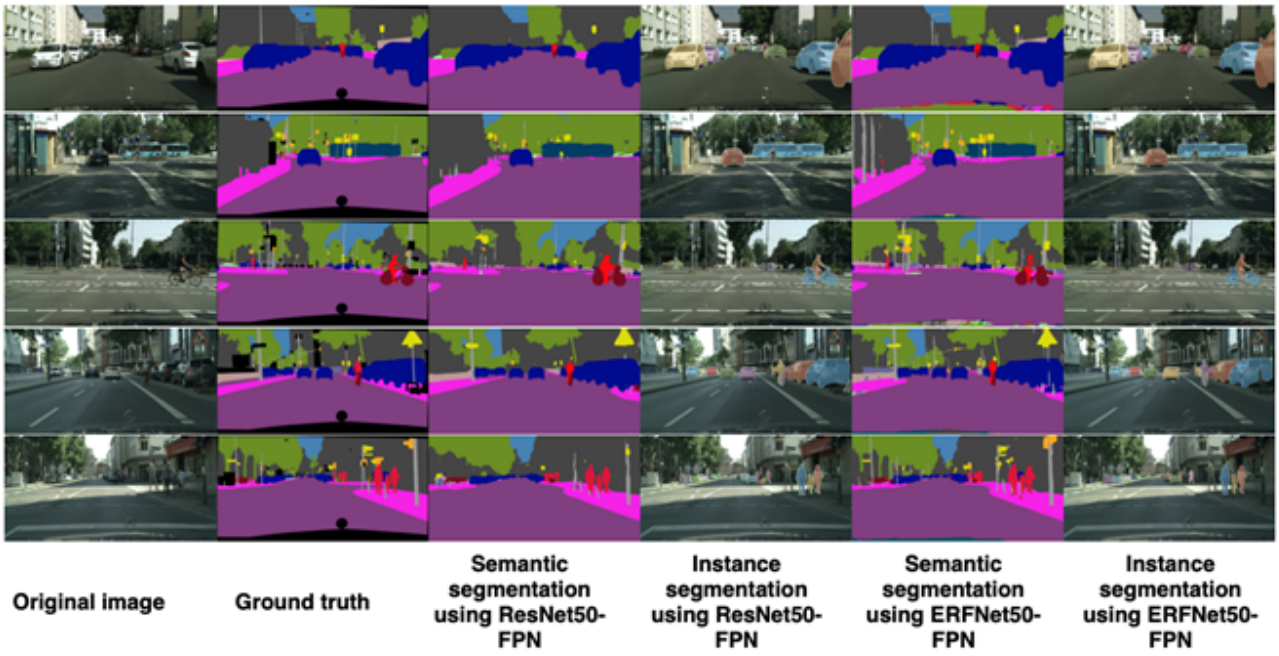

Fig. T2.2.2 Results from the paper “Efficient instance and semantic segmentation for automated driving”. The ResNet50-FPN is the baseline network and the ERFNet50-FPN is the proposed network. Compared to the baseline, the proposed solution increased the segmentation mIoU with 4.5%.

Reference

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

A3. Semantic Optimal Control

Task 3.2. Representations and Methods for Efficient Computation

Existing state-of-the-art estimation systems can detect 2d poses of multiple people in images quite reliably. In contrast, 3d pose estimation from a single image is ill-posed due to occlusion and depth ambiguities. Assuming access to multiple cameras, or given an active system able to position itself to observe the scene from multiple viewpoints, reconstructing 3d pose from 2d measurements becomes well-posed within the framework of standard multi-view geometry. Less clear is what is an informative set of viewpoints for accurate 3d reconstruction, particularly in complex scenes, where people are occluded by others or by scene objects. In order to address the view selection problem in a principled way, we here introduce ACTOR, an active triangulation agent for 3d human pose reconstruction. Our fully trainable agent consists of a 2d pose estimation network (any of which would work) and a deep reinforcement learning-based policy for camera viewpoint selection. The policy predicts observation viewpoints, the number of which varies adaptively depending on scene content, and the associated images are fed

to an underlying pose estimator. Importantly, training the policy requires no annotations - given a 2d pose estimator, ACTOR is trained in a self-supervised manner. In extensive evaluations on complex multi-people scenes filmed in a Panoptic dome, under multiple viewpoints, we compare our active triangulation agent to strong multi-view baselines, and show that ACTOR produces significantly more accurate 3d pose reconstructions. We also provide a proof-of-concept experiment indicating the potential of connecting our view selection policy to a physical drone observer. For qualitative results, please have a look at fig. T3.2.3 below.

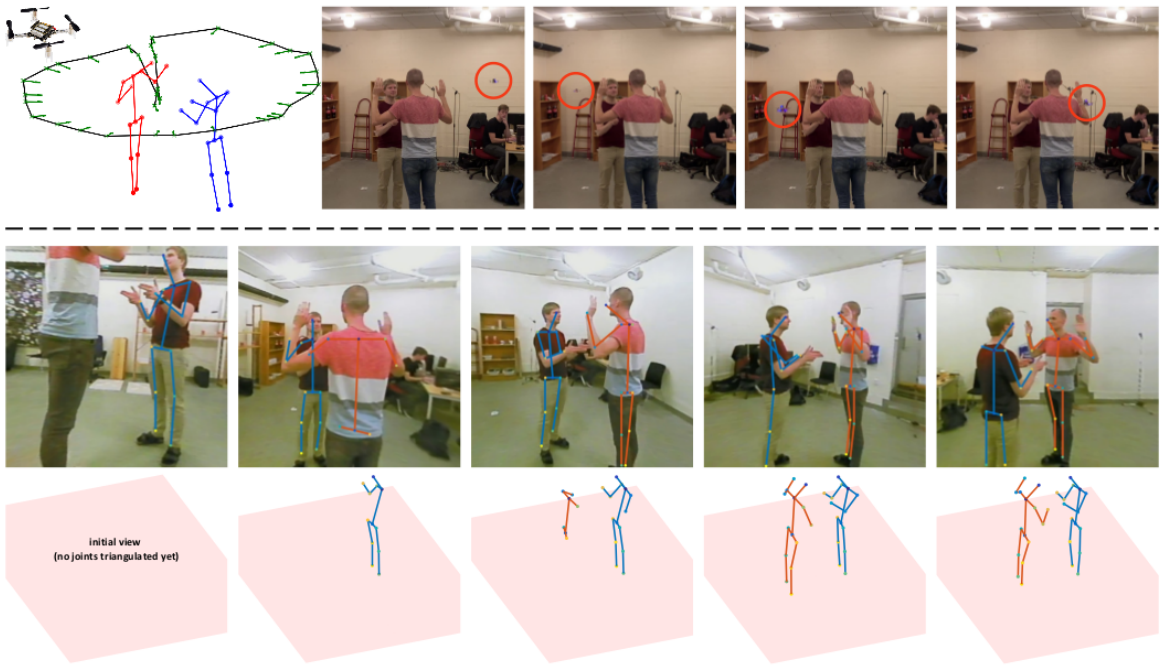

Fig. T3.2.1 From domes to drones. Proof-of-concept experiment illustrating that ACTOR can be connected to an active drone observer to reconstruct 3d poses from informative viewpoints. Above the dashed line to the left we show the drone's loop (the sharp peak is due to take-off and landing), with sampled camera locations as green arrows. We also show the 3d pose reconstructions obtained by triangulating from all 33 sampled camera locations. The 9-by-9 cm Crazyflie drone used is shown in the very top left corner; it can be used safely due to its small size and weight. Sample locations of the drone are also shown above the line (drone locations are highlighted with red circles in images). Below the line we show views seen by ACTOR and aggregated 3d pose reconstructions. After observing 5 viewpoints, the two bodies are fully 3d reconstructed, with an average 2d reprojection error of 11.5 pixels (averaged over all 33 cameras), significantly better than the exhaustively triangulated reconstructions to the left, with an average reprojection error of 35.4 pixels.

Reference

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