U3D-MOLTS: Unified 3D Monocular Object Localization, Tracking and Segmentation

Haotian Zhang, Yizhou Wang, Zhongyu Jiang, Cheng-Yen Yang, Jie Mei, Jiarui Cai, Jenq-Neng Hwang
University of Washington, Seattle, WA, USA
{haotiz, ywang26, zyjiang, cycyang, jiemei, jrcai, hwang}@uw.edu

Kwang-Ju Kim, Pyong-Kun Kim
Electronics and Telecommunications Research Institute (ETRI), South Korea
{kwangju, iros}@etri.re.kr

Abstract

Sensing and perception systems for autonomous driving vehicles in road scenes are composed of four crucial components: object detection, tracking, panoptic segmentation, and 3D localization. While all these components are inherently intertwined, most relevant papers tend to only focus on a subset of these components. We separate the unified video panoptic segmentation tasks into two subtasks: 1) multi-object tracking and segmentation, and 2) semantic segmentation. For the first stage, we propose a unified monocular 3D based framework that effectively tracks detected moving objects over time and estimates their 3D localization information as well as instance segmentation masks from a sequence of 2D images captured from a dash camera on a moving vehicle. Our system contains an RCNN-based Localization for Tracking Network (Loc4Trk-Net). The object association leverages deep pairwise contrastive learning to identify objects in various poses and viewpoints with appearance cues. A straightforward combination of a 3D Kalman filter and the Hungarian algorithm is further utilized for robust instance association via both feature similarity and 3D localization information. For the second stage, we adopt the existing DeepLabV3+ for semantic segmentation and further enhanced the performance with data augmentation using label propagation. Our proposed pipeline achieves an STQ score of 67.55 on the KITTI-STEP dataset as well as the state-of-the-art performance on the KITTI-MOTS leaderboard.

1. Introduction

3D Monocular object localization, tracking, and segmentation are inherently ill-posed. The 3D detection method is challenging in the absence of depth measurements or strong priors given a single image, which often requires a large amount of training data and is hard to adapt since they are sensitive to training data. To overcome these problems, our proposed unified monocular 3D framework begins with an easy-to-train RCNN-based Localization for Tracking Network (Loc4Trk-Net), which is only trained with limited amounts of training data, not only to generate 2D bounding boxes and instance masks but also to provide proper initialization of detected objects' 3D orientations and distances; Frame-by-frame detections are never perfect. Given a strong localization basis, short-term 3D tracking tends to be more robust and long-term 3D tracking becomes possible. At the same time, 3D tracking information across multiple frames can further assist 3D localization as well by recovering missing/unreliable detections. In addition, self-supervised spatial attention is also applied to our model to learn an instance-aware embedding for each object, which is an instance descriptor represented as a vector in a latent space via deep contrastive learning. Robust tracking results are obtained by associating the detections with the learned features and their historical trajectories using an online 3D Kalman filter and Hungarian matching algorithm. In summary, the proposed method claims the following contributions:

- An RCNN-based Loc4Trk-Net is proposed to not only generate 2D bounding box and instance masks but also simultaneously predict both the 3D orientation and distance of vehicles in the camera coordinate.
- Instance specific features, which are learned jointly with the detection task, utilize the instance masks as spatial attention to serve as a better embedding feature for tracking association.
- 3D object tracking, which uses the learned instance-aware feature via pairwise contrastive learning, is incorporated. A straightforward combination of a 3D

Kalman filter and the Hungarian algorithm is used for online state estimation and robust data association.

- DeepLabV3+ [1] is adopted here for semantic segmentation. To further improve its accuracy, we apply a novel data augmentation method [13] by generating pseudo images and labels.
- The proposed framework achieves an STQ of 67.55 on the KITTI-STEP leaderboard. Apart from that, we apply our model to experiment on the KITTI-MOTS dataset, and it also achieves the state-of-the-art performance on the KITTI-MOTS dataset.

2. Related Work

Multi-Object Tracking. Recent multiple object tracking (MOT) methods have largely employed tracking-by-detection schemes, meaning that tracking is done through the association of detected objects across time. Most works [12] on MOT are typically done in 2D image space. However, lack of depth information in 2D tracking causes failure in tracking objects long-term due to disappearances or occlusions. Given LiDAR point cloud, [8] uses standard 3D Kalman filters and Hungarian algorithms to associate detections from LiDAR, resulting in fewer ID switches and can perform long-term tracking. Yet LiDAR has its own drawbacks, such as high cost and sensitivity to adverse weather conditions. These limitations suggest that employing a LiDAR-based object tracking system is unrealistic in practical, day-to-day applications. Our work is in line with the recent developments in 3D object tracking fields and aims to improve data association by leveraging 3D information, but goes beyond this by integrating both visual cues and 3D localization information using only a monocular camera without additional sensors to track objects in 3D.

Multi-Object Tracking and Segmentation. MOTS is proposed as a new task to track multiple objects with instance segmentation. Voigtlaende et al. [6] propose a baseline approach, Track R-CNN, which can jointly address detection, tracking, and segmentation via a single convolutional network. While the aforementioned method is able to produce tracking outputs with segmentation masks, the network is trained under multiple task, resulting in increasing the tracking performance while degrading the detection and segmentation performance.

Video Panoptic Segmentation. Recently, the above multi-object tracking with instance segmentation has also been elevated to the panoptic domain. The new task requires not only generating the tracking IDs along with instance segmentation results across video frames but also do the semantic segmentation at the same time. [7] proposes a new benchmark dataset named KITTI-STEP addressing the long-term segmentation and tracking problem, and together with a new evaluation metric, which provides an important insight towards a denser, pixel-precise video understanding.

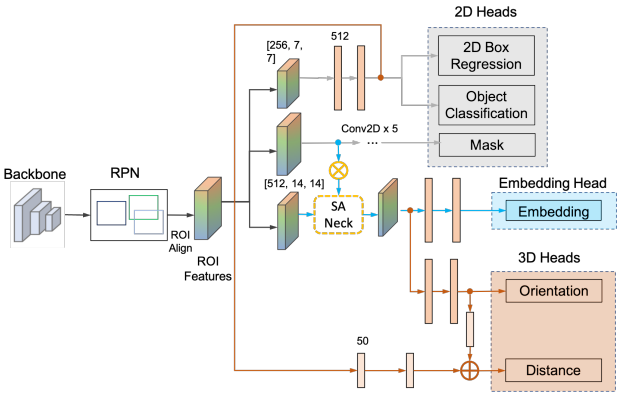

Figure 1. **Detailed Architecture for Loc4Trk-Net.** The upper two branches (in gray) are the typical Mask-RCNN detection framework. The middle branch (blue) is the embedding head, with the help of spatial attention (SA) neck (green), which heavily weighs on the foreground object to enhance instance-specific appearance features and suppress the noise in the background. The bottom two branches predict the 3D orientation and distance in the camera coordinates of the detected object (brown).

3. The Proposed Method

3.1. Multi-Object Tracking and Segmentation

Network Architectures. The proposed Localization for Tracking Network (Loc4Trk-Net) is built upon a canonical two-stage object detection network, Mask R-CNN [2]. Loc4Trk-Net augments the Mask R-CNN model with additional learning objectives. The first stage of the network is a 2D object detection network, which extracts and scores region proposals by RPN, followed by the ROIAlign for feature cropping. Based on the top-scoring proposals, a convolutional encoder is used to refine the cropped features, which are then fed into five separate heads. The second stage of the network consists of both classical and customized heads. For the 2D part we use three heads for standard multi-class classification, 2D box refinement and mask generation respectively. Two more heads are introduced to predict object 3D orientation and distance in the camera coordinates. One additional embedding head is introduced to train a discriminative feature embeddings to associate detections and tracklets. The detailed architecture is shown in Fig. 1.

Spatial Attention (SA) Neck. The intuition of the spatial attention (SA) neck is to highlight the foreground (target of interest) and suppress the background, so that more concentrated appearance features can be obtained. Details of the SA neck is shown in Fig. 1 (green), where ROI features are pooled and flattened for classification and bounding box regression. Simultaneously, they are passed through four 2D convolutional layers and a pixel-wise sigmoid operation

to generate the SA map, indicating the probability of objectiveness. With the intermediate output of the SA operation, several 2D convolution layers with kernel size 1 are applied to produce the object mask. Meanwhile, the ROI features are multiplied by the SA map to purify pixels which belong to the target and single-dimensional feature is further extracted by fully connected layers.

Instance Embedding Head. The multi-object tracking problem requires distinguishable feature embeddings to match detections and tracklets. We use RPN to generate RoIs from the two images and RoIAlign to obtain their feature maps from different levels in the feature pyramid network (FPN) according to their scales. An extra lightweight embedding head is added to extract features for each RoI, as shown in Fig. 1 (blue). An RoI is defined as positive to an object if they have an IoU higher than 0.7, or negative if they have an IoU lower than 0.3 in our settings. The matching of RoIs on two frames is positive if the two regions are associated with the same object and negative otherwise. By balancing positive and negative samples, we encourage the embedding head to learn a feature embedding that can effectively discriminate between instances, while being invariant to perturbations like changes in viewpoint or lightning.

3D Orientation and Distance Head. The orientation head takes the features from SA neck as input to generate the 3D orientation output. The distance head takes a concatenated input, from both depth-aware ROIAligned feature maps ($256\times14\times14$) and convolved 512-dim features for bounding-box classification/regression, to form more informative input features for 3D distance, which is demonstrated in Fig. 1 (brown). The concatenated features are assumed to implicitly encode the 3D orientation information and pre-defined object size information via the incorporation of the convolved 512-dim features. The outputs from Loc4Trk-Net are particularly crucial for tracking and segmentation tasks, where the feature embedding, 3D information as well as their 2D instance masks of the objects are of primary importance.

Data Association and Tracking. For simple design and real-time efficiency, we use a conventional way to solve the association between the predicted 3D Kalman states and newly arrived measurements, which is to build assignment problems that can be solved using the Hungarian algorithm. To incorporate motion information, the detectionsvand predicted trajectories are associated using the Hungarian algorithm. An affinity matrix is constructed by computing the 3D Intersection of Union (IoU) or negative center distance between every pair of the trajectory and detection. To incoporate the appearance information, our second metric measures the smallest *cosine* distance between the i-th track and j-th detection in appearance space. We combine both metrics to get matched trajectories and detections using a weighted sum following [9]. Based on the tracking results, we are not only able to associate every object across frames, but also can deal with errors caused by the occlusions and missing detections. For those missing detections, we use Huber regression for detection interpolations.

3.2. Semantic Segmentation

We adopted the DeepLabV3+ [1], which is the state-of-the-art method in KITTI semantic dataset for semantic segmentation. To improve semantic segmentation results, Atrous Convolution is utilized in DeepLabV3+, for integrating global and local features for the network. Furthermore, Zhu et al. [13] introduce a simple yet efficient data augmentation pipeline for improving Semantic Segmentation training. They take the image I_t and label L_t as reference jointly and predict images $I_{t\pm s}$ and labels $L_{t\pm s}$ for data augmentation. As a result, the dataset can be scaled by a factor $2k+1$. Besides that, Boundary Label Relaxation is introduced for better object semantic boundary estimation. Thus, by combining the above instance results from 3.1 and semantic segmentation methods, we achieve segmentation quality SQ of 64.04 in the KITTI-STEP.

4. Experiment Results

4.1. Dataset and Evaluation Metrics

KITTI-STEP [7] is a driving scenario dataset for both car and pedestrian tracking task. It consists of 21 training sequences and 29 testing sequences. The evaluation metrics is the segmentation and tracking quality (STQ) consisting of two factors, association quality (AQ) and segmentation quality (SQ), that measure the tracking and segmentation quality respectively. KITTI-MOTS [6] has the same train and test sequences, and we evaluate our performance using HOTA metrics [4], which accumulates the soft number of true positives, false positives, and ID switches.

4.2. KITTI-STEP Performance

The performance of KITTI-STEP using STQ, which measures segmentation as well as detection and tracking quality. Our method currently ranks the first place among the total valid submissions. The performance of top-selected algorithms among all competitors is shown in Table 1.

4.3. KITTI-MOTS Performance

Our performance on multi-object tracking and segmentation in the KITTI-MOTS is also shown in Table. 2. Upon the time of submission, we are the 1^{st} place among all the image-based methods. Vip-DeepLab [5] tries to approach the task by jointly performing monocular depth estimation and video panoptic segmentation, though they require additional ground-truth for training the depth esimation module. ReMOTS [11] proposed an intra-frame self-supervised

Table 1. Competition results on KITTI-STEP *test* set, ours is marked **bold**.

Method	**STQ** ↑	AQ ↑	SQ(IoU) ↑
Motion-DeepLab [7]	52.19	45.55	59.81
HybridTracker	54.99	54.44	55.54
siain	57.87	55.16	60.71
EffPS_MM	62.93	61.49	64.41
REPEAT	67.13	65.81	**68.49**
IPL_ETRI (Ours)	**67.55**	**71.26**	64.04

Table 2. Performance of multi-object tracking and segmentation methods on KITTI-MOTS *test* set, ours is marked **bold**.

Tracker	**HOTA** ↑	DetA ↑	AssA ↑
ViP-DeepLab [5]	76.38	**82.70**	70.93
ReMOTS [11]	71.61	78.32	65.98
PointTrackV2 [10]	66.33	83.12	53.38
TrackR-CNN [6]	56.63	69.90	46.53
MOTSFusion [3]	73.63	75.44	72.39
IPL_ETRI (Ours)	**79.57**	79.66	**80.00**

triplet construction network to learn mask features for both training and testing set for Re-ID. PointTrackV2 [10] distinguish the foreground and background by regarding the object's mask and its surrounding environment as two sets of 2D point clouds. However, these two methods are assuming the accurate initialization of segmentation from a pretrained optical-flow estimation network.

5. Conclusion

We propose an unified 3D monocular localization, tracking and segmentation pipeline, combining with pairwise contrastive learning and 3D instance estimation, to tracking moving vehicles in a 3D world. Our proposed pipeline consists of four parts: an RCNN-based Localization for Tracking Network (Loc4Trk-Net), cross-frames contrastive feature learning modules, and a simple but effective 3D Kalman filter. DeepLabV3+ is adopted here for semantic segmentation. To further improve its accuracy, we apply a novel data augmentation method by generating pseudo images and labels. Extensive experiments and ablation studies have shown our method is effective and robust under different autonomous driving scenarios. Overall, our method currently ranks the 1^{st} place on both KITTI-STEP and KITTI-MOTS leaderboard.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 3

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[3] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *IEEE Robotics and Automation Letters*, 5(2):1803–1810, 2020. 4

[4] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, pages 1–31, 2020. 3

[5] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. *arXiv preprint arXiv:2012.05258*, 2020. 3, 4

[6] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7942–7951, 2019. 2, 3, 4

[7] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021. 2, 3, 4

[8] Xinshuo Weng and Kris Kitani. A baseline for 3d multi-object tracking. *arXiv preprint arXiv:1907.03961*, 2019. 2

[9] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 3

[10] Zhenbo Xu, Wei Zhang, Xiao Tan, Wei Yang, Xiangbo Su, Yuchen Yuan, Hongwu Zhang, Shilei Wen, Errui Ding, and Liusheng Huang. Pointtrack++ for effective online multi-object tracking and segmentation. *arXiv preprint arXiv:2007.01549*, 2020. 4

[11] Fan Yang, Xin Chang, Chenyu Dang, Ziqiang Zheng, Sakriani Sakti, Satoshi Nakamura, and Yang Wu. Remots: Self-supervised refining multi-object tracking and segmentation. *arXiv e-prints*, pages arXiv–2007, 2020. 3, 4

[12] Zhimeng Zhang, Jianan Wu, Xuan Zhang, and Chi Zhang. Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. *arXiv preprint arXiv:1712.09531*, 2017. 2

[13] Yi Zhu, Karan Sapra, Fitsum A Reda, Kevin J Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8856–8865, 2019. 2, 3