LIFTS: Lidar and Monocular Image Fusion for Multi-Object Tracking and Segmentation

Haotian Zhang, Yizhou Wang, Jiarui Cai, Hung-Min Hsu, Haorui Ji, Jenq-Neng Hwang
Department of Electrical and Computer Engineering
University of Washington, Seattle, WA, USA
{haotiz, ywang26, jrcai, hmhsu, hji2, hwang}@uw.edu

Abstract

In recent years, the computer vision society has made significant progress in multi-object tracking (MOT) and video object segmentation (VOS) respectively. Further progress can be achieved by effectively combining the following tasks together – detection, segmentation and tracking. In this work, we propose a multi-stage framework called "Lidar and monocular Image Fusion based multi-object Tracking and Segmentation (LIFTS)" for multi-object tracking and segmentation (MOTS). In the first stage, we use a 3D Part-Aware and Aggregation Network detector on the point cloud data to get 3D object locations. Then a graph-based 3D TrackletNet Tracker (3D TNT), which takes both CNN appearance features and object spatial information of detections, is applied to robustly associate objects along time. The second stage involves a Cascade Mask R-CNN based network with PointRend head for obtaining instance segmentation results from monocular images. Its input pre-computed region proposals are generated from projecting 3D detections in the first stage onto a 2D image plane. Moreover, two post-processing techniques are further applied in the last stage: (1) generated mask results are refined by an optical-flow guided instance segmentation network; (2) object re-identification (ReID) is applied to recover ID switches caused by long-term occlusion; Overall, our proposed framework is evaluated on BMTT Challenge 2020 Track2: KITTI-MOTS dataset and achieves a 79.6 sMOTSA for Car and 64.9 for Pedestrian, with the 2^{nd} place ranking in the competition.

1. Introduction

Multi-Object Tracking (MOT) is the task of associating objects in a video and assigning consistent IDs for the same identities. On the other hand, Video Object Segmentation (VOS) task is aimed to generate the object segmentation masks in video frames. However, many available approaches can successfully track objects when they are consistently visible, but fail in the long-term through either disappearances or occlusions, which are caused by heterogeneous objects, interacting objects, edge ambiguous and shape complexity. Jointly solving the two problems of multi-object tracking and segmentation (MOTS) can overcome respective difficulties and improve both of their performances.

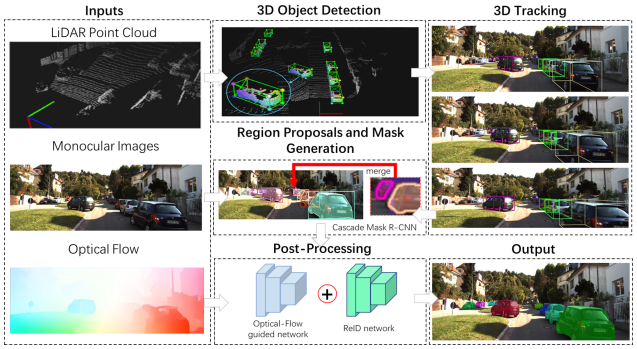

Figure 1. The framework of the proposed LIFTS, which consists of a three-stage pipeline: 3D object detection and tracking, pre-computed proposals and masks generation, and post-processing.

To tackle the MOTS problems, we propose a multi-stage framework named "Lidar and monocular Image Fusion for multi-object Tracking and Segmentation (LIFTS)". The LIFTS consists of a three-stage pipeline, which is shown in Figure 1. First, given the LiDAR point cloud, a 3D Part-Aware and Aggregation Network is adopted to get accurate 3D object locations. A graph-based TrackletNet, which takes both CNN appearance and object spatial information, is then applied to robustly associate objects to form tracklets along time. We additionally interpolate and recover the unreliable/missing detections in each tracklet to form longer trajectories. In the second stage, we project each frame of these trajectories onto 2D image plane using the camera intrinsics and treat them as the pre-computed region

proposals for a Cascade Mask R-CNN [2] based network with a PointRend [4] segmentation branch. For those objects that are not detected by the 3D detector from LiDAR point cloud data but detected by the 2D detector from images, we use Hungarian algorithm to merge these detections with existing trajectories. This is able to bridge 3D world space and 2D image space consistently and produces a high-quality mask for each instance in every frame. Moreover, two post-processing stages are further applied in the last stage: (1) the generated masks from the second stage are refined by a proposed optical-flow guided instance segmentation network; (2) the refined masks are then used for object re-identification (ReID) to recover ID switches caused by long-term occlusions.

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 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 and occlusions. Given LiDAR point cloud, [11] uses standard 3D Kalman filters and Hungarian algorithms to associate detections from LiDAR, which causes fewer ID switches and can perform long-term tracking.

Multi-Object Tracking and Segmentation. MOTS is proposed as a new task to track multiple objects with instance segmentation. Voigtlaende et al. [9] 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.

3. The Proposed Method

3.1. 3D Object Detection and Tracking

3D Part-Aware and Aggregation Network. We adopt a state-of-the-art point-cloud-based 3D object detector, the part-aware and aggregation neural network (Part-A^2 net) [8]. The Part-A^2 detector produces 3D bounding boxes parameterized with $(x, y, z, h, w, l, \theta)$, where (x, y, z) are the box center coordinates, (h, w, l) are the height, width and length of each box respectively, and θ is the orientation angle of each box from the bird's eye view.

3D TrackletNet Tracker. To take advantage of the temporal consistency for improving the localization performance further, we need tracking to associate corresponding detected objects along time. The proposed 3D TrackletNet Tracker (3D TNT) takes both discriminative CNN appearance features and accurate object spatial information from each frame to ensure tracking robustness. The 3D TNT is extended from the 2D TNT [10], which builds a graph-based model that takes 2D tracklets as the vertices and use a multi-scale CNN network to measure the connectivity between two tracklets. Our 3D TrackletNet Tracker consist of three key components:

(i) *Tracklet Generation*: Given the refined object localization of each frame, generated by 2D box appearance similarity based on CNN features derived from the FaceNet [7] and 3D intersection-over-union (3D IoU) between adjacent frames, is denoted as a node ($v \in V$) in the graph.

(ii) *Connectivity Measurement*: Between every two tracklets, the connectivity (similarity) $p_e(e \in E)$ is measured as the edge weight in the graph model. To calculate the connectivity, a multi-scale TrackletNet is built as a classifier, which can concatenate both temporal (multi-frame) and appearance features for the likelihood estimation. For each frame t, a vector consisting of the 7-D detected object measurements $(x, y, z, h, w, l, \theta)$ from the Part-A^2 detector, concatenated by an 512-D embedding appearance feature extracted from the FaceNet, is used to represent an individual feature of the input frame.

(iii) *Graph-based Clustering*: After the tracklet graph is built, graph partition and clustering techniques, i.e., assign, merge, split, switch, and break operations are iteratively performed to minimize the total cost on the whole graph.

Based on the tracking results from the 3D TNT, 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 unreliable/missing detections, we use Lagrangian interpolation to recover/fill-in those frames to form longer trajectories.

3.2. Pre-Computed Proposals and Masks Generation

In Sec. 3.1, accurate locations are obtained by a 3D object detector and objects are robustly associated across frames using the proposed 3D TrackletNet Tracker. In order to produce outputs with segmentation masks, we project all the 3D bounding boxes inside each frame onto the 2D image plane by using camera intrinsics, object locations and orientation angles and treat the projected ROIs as the pre-computed region proposals to a two-stage image-based object detection framework. For smaller objects that are not detected by Part-A^2 network but are detected by the image-based detector, we use Hungarian algorithm to merge these detections with existing trajectories.

We utilize the Cascade Mask R-CNN framework as our basic architecture. Each detector inside the cascade is sequentially more effective in selecting higher quality detections compared with its predecessor. The network can then

Figure 2. Qualitative results of the proposed LIFTS method on KITTI-MOTS datasets. The top two rows are the results for Cars, and the bottom two rows show the results for Pedestrians.

make itself capable of handling proposals through multiple quality levels and generate better detection results.

In addition to the detection head, we use PointRend [4] as our mask head for segmentation. Compared with standard mask head, PointRent head, which iteratively renders the output mask in a coarse-to-fine fashion, upsamples its previously predicted segmentation using bilinear interpolation and then selects 50 most uncertain points to predict their labels using a point-wise feature representation. In this case it can predict masks with substantially finer details around object boundaries.

3.3. Post-Processing

Optical-Flow Guided Mask Propagation Network. Followed the idea by [1], which shows the highly accurate object segmentation in videos can be achieved by using temporal information, the segmentation mask can be further refined from the previous frame's "rough" estimate of a tracked object. In order to guide the pixel labeling network to segment the object of interest, we begin by expanding the network input from `RGB` to `RGB+mask` channel. The extra mask channel is meant to provide an estimate of the visible area of the object in the current frame, its approximate location and shape. Given an initial mask estimate from the previous frame $t-1$ in Sec. 3.2, we train the network to provide a refined mask output for the current frame t.

We also consider to employ the optical flow as a source of additional information to guide the segmentation. More specifically, given a video sequence, we compute the optical flow using FlowNet2 [3]. In parallel to the above framework, we proceed to compute a second output mask using the magnitude of the optical flow field as the input image. We then fuse by averaging the output scores given by the two parallel networks. It can be shown in the experimental results, optical flow provides complementary information to the mask quality, improving the overall performance.

Re-identification Network. A ReID approach is also applied to reconnect tracklets due to occlusions or abrupt motions. We use the trajectory-level of features for ReID in the tracking refinement process. For frame-level feature extraction, we adopt the ResNet50 network pre-trained on ImageNet as our feature extractor. Furthermore, temporal information is also considered to establish a more representative feature by using temporal attention (TA) to convert weighted average of the frame-level features and into clip-level features. Note that some frames of the object might be highly occluded by other objects, and we try to lower the weights of these frames. Finally, we add a max-pooling layer for these clip-level features to generate the final tracklet-level feature.

4. Experiments

4.1. Dataset and Evaluation

KITTI-MOTS [9] is a driving scenario dataset for both car and pedestrian tracking task. It consists of 21 training sequences and 29 testing sequences. We evaluate our performance based on sMOTSA metrics, which accumulates the soft number of true positives, false positives, and ID switches.

4.2. KITTI-MOTS Performance

The performance of MOTS for both Car and Pedestrian are evaluated using sMOTA, which measures segmentation as well as detection and tracking quality. Qualitative performance is shown in Fig. 2. In the BMTT Challenge 2020 Track2 (KITTI-MOTS), our method ranks the second place among the total 16 valid submissions. The performance of top-selected algorithms is shown in Table 1 for Car and Table 2 for Pedestrian.

Method	sMOTA ↑	MOTSA	MOTSP	MOTSAL	MODSA	MODSP	MT	ML	IDS ↓	Frag
PointTrack++	82.80	92.60	89.70	93.30	93.30	92.10	89.10	1.2	270	584
MCFPA	77.00	87.70	88.30	89.10	89.10	90.80	82.90	0.6	503	724
GMPHD_SAF	75.40	86.70	87.50	88.20	88.20	90.10	82.00	0.6	549	874
MOTSFusion [6]	75.00	84.10	89.30	84.70	84.70	91.70	66.10	6.2	201	572
TrackR-CNN [9]	67.00	79.60	85.10	81.50	81.50	88.30	74.90	2.3	692	1058
LIFTS (ours)	79.60	89.60	89.00	89.90	89.90	91.40	79.10	2.9	114	532

Table 1. Competition results for **Car** of KITTI-MOTS, ours is marked **bold**.

Method	sMOTA ↑	MOTSA	MOTSP	MOTSAL	MODSA	MODSP	MT	ML	IDS ↓	Frag
PointTrack++	68.10	83.60	82.20	84.80	84.80	94.00	66.70	4.80	250	521
MCFPA	67.20	83.00	81.90	84.30	84.30	93.80	67.00	3.00	265	484
GMPHD_SAF	62.80	78.20	81.60	80.40	80.50	93.70	59.30	4.80	474	696
MOTSFusion [5]	58.70	72.90	81.50	74.20	74.20	94.10	47.40	15.60	279	534
TrackR-CNN [9]	47.30	66.10	74.60	68.40	68.40	91.80	45.60	13.30	481	861
LIFTS (ours)	64.90	80.90	81.00	81.90	81.90	93.60	61.50	8.90	206	577

Table 2. Competition results for **Pedestrian** of KITTI-MOTS, ours is marked **bold**.

5. Conclusion

We have presented a framework in which both tracking and segmentation can be performed together and can benefit from each other. We first use a Part-Aware and Aggregation Network given the LiDAR point cloud data to get accurate 3D object locations, then a proposed graph-based 3D TrackletNet Tracker is applied to associate object across frames. We treat the projected 2D ROI as the pre-computed region proposals and send them into a cascade Mask R-CNN network with PointRend segmentation. Finally, a proposed optical-flow guided instance segmentation network and a ReID approach is applied to further refine both segmentation and tracking results. Quantitative and qualitative experiments have demonstrated that our system can achieve high accuracy in sMOTA for both Cars and Pedestrians and outperforms other competing methods.

References

[1] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9739–9748, 2020. 3

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2

[3] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 3

[4] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. *arXiv preprint arXiv:1912.08193*, 2019. 2, 3

[5] 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

[6] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580. Springer, 2018. 4

[7] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2

[8] Shaoshuai Shi, Zhe Wang, Xiaogang Wang, and Hongsheng Li. Part-aˆ 2 net: 3d part-aware and aggregation neural network for object detection from point cloud. *arXiv preprint arXiv:1907.03670*, 2019. 2

[9] 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

[10] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 482–490, 2019. 2

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

[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