REAL-TIME DEEP NEURAL NETWORKS FOR MULTIPLE OBJECT TRACKING AND SEGMENTATION ON MONOCULAR VIDEO

I. Basharov[1], D. Yudin [1, *]

[1] Intelligent Transport Lab., Moscow Institute of Physics and Technology, Dologoprudny, Russia
yudin.da@mipt.ru

KEY WORDS: Multiple object tracking, Instance segmentation, Deep neural network, Real time, Monocular video

ABSTRACT:
The paper is devoted to the task of multiple objects tracking and segmentation on monocular video, which was obtained by the camera of unmanned ground vehicle. The authors investigate various architectures of deep neural networks for this task solution. Special attention is paid to deep models providing inference in real time. The authors proposed an approach based on combining the modern SOLOv2 instance segmentation model, a neural network model for embedding generation for each found object, and a modified Hungarian tracking algorithm. The Hungarian algorithm was modified taking into account the geometric constraints on the positions of the found objects on the sequence of images. The investigated solution is a development and improvement of the state-of-the-art PointTrack method. The effectiveness of the proposed approach is demonstrated quantitatively and qualitatively on the popular KITTI MOTS dataset collected using the cameras of a driverless car. The software implementation of the approach was carried out. The acceleration of the procedure for the formation of a two-dimensional point cloud in the found image segment was done using the NVidia CUDA technology. At the same time, the proposed instance segmentation module provides a mean processing time of one image of 68 ms, the embedding and tracking module of 24 ms using the NVidia Tesla V100 GPU. This indicates that the proposed solution is promising for on-board computer vision systems for both unmanned vehicles and various robotic platforms.

1. INTRODUCTION

Multiple object tracking (MOT) task is very important for a large number of applications. By solving it, we can build trajectories for dynamic obstacles surrounding robotic platforms or unmanned vehicles and improve their safety. In the general case, tracking is based on the results of object recognition in the frame (Yudin et al, 2019). All methods can be divided by online (Bewley et al., 2016, Bochinski et al., 2017, Payer et al., 2018) using current and previous frames, and offline (Voigtlaender et al., 2019, Luiten et al., 2018) using future frames features.

Our work is devoted to the study of online tracking objects methods. The purpose of work is to implement the model in intelligent transport systems. A distinctive feature of work in such conditions is the high quality and speed of image processing. Meaning that image recognition methods in 2D are often faster than those in 3D point clouds, we chose instance segmentation on monocular video for object tracking.

The approach developed in this paper contains the following contributions:

- improvements of PointTrack method (Xu et al., 2020) were proposed, which consist in replacing the basic instance segmentation model with the high-speed SOLOv2 model (Wang et al., 2020) and modifying the model that creates embedding for each a found object, taking into account its category;
- modification of the Hungarian algorithm was made taking into account a geometric constraint of the found objects on an image sequence;
- the software implementation of the approach was carried out, including the procedure acceleration for a two-dimensional point cloud formation in a found image segment using the NVidia CUDA technology.

2. RELATED WORK

Online object tracking methods assume that detection results are available; it focuses on the data association. SORT (Bewley et al., 2016) uses the Kalman filter to predict future object locations, and the Hungarian algorithm (Kuhn, 1955) to match object representations. The IoU tracker (Bochinski et al., 2017) directly links instances of nearby frames by their spatial overlap.

Offline methods use an internal segment allocation. ATOM (Danelljan et al., 2019) not only uses object overlap, but additionally classifies to obtain multiclass tracking. FAMNet (Chu et al., 2019) jointly optimizes feature extraction and similarity estimation.

Unlike 2D bounding boxes which might overlap heavily in crowded scenes, per-pixel segments locate objects precisely (Staroverov et al., 2020). This idea served to develop of MOTS direction – multiple objects tracking and segmentation. Track R-CNN method (Voigtlaender et al., 2019) uses 3D convolutions to obtain temporal information. MOTSNet (Porzi et al., 2020) proposed a new mask pooling layer to improve object association. MOTSFusion offline method (Luiten et al., 2020) improves the quality of tracking by combining the results of determining the depth map, optical flow, odometry, as well as the reconstructed surrounding 3D scene.

Also using LiDAR point cloud authors of LIFTS (Zhang et al, 2019) do tracking in 3D scenes. 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. Second, they projected each frame of these trajectories into 2D image plane using the camera intrinsic matrix and treated them as the pre-computed region proposals for a Cascade Mask R-CNN based network with a PointRend segmentation branch. They used Hungarian algorithm to merge LiDAR and Image detections with existing trajectories.

The monocular image-based state-of-the-art approach PointTrack (Xu et al., 2020) should be noted individually, which forms an object vector representation using a 2D point cloud. It demonstrates the high quality of indexing objects in real-time on the data of the KITTI MOTS competition (Voigtlaender et al., 2019). An improvement of PointTrack is PointTrack ++ (Zhenbo et al., 2020). Firstly, in the instance segmentation stage, the

* Corresponding author

methods adopted a semantic segmentation decoder trained with focal loss to improve the instance selection quality. Secondly, to further boost the segmentation performance, a data augmentation strategy is proposed by copy-and-paste instances into training images.

STE (Hu et al., 2019) introduces a new spatial-temporal embedding loss to generate temporally consistent instance segmentation and regard the mean embeddings of all pixels on segments as the instance embedding for data association.

Tracklet-Plane Matching (Yuchen et al, 2019) focused on the tracking task itself with pre-defined detection and segmentation results in the scene, and propose a framework that integrates segmentation-based feature extraction, short tracklet construction, and tracklet-plane matching for long trajectories.

The next method based on tracklet is ReMOTS (Yang et al., 2020). It takes four steps to refine MOTS results from the data association perspective. Firstly, training the appearance encoder using predicted masks. Secondly, associating observations across adjacent frames to form short-term tracklets. After that, training the appearance encoder using short-term tracklets as reliable pseudo labels. Further merging short-term tracklets to long-term tracklets utilizing adopted appearance features and thresholds.

Authors of EagerMOT (Kim et al, 2020) proposed a tracking formulation that eagerly integrates all available object observations from both sensor modalities in order to obtain a well-informed interpretation of the scene dynamics. This method presents a framework that combines the localization accuracy of 3D-based detectors with the precision of 2D object detectors.

Authors of IA-MOT (Cai et al., 2020) proposed a tracking framework that can track multiple objects in either static or moving cameras by jointly considering the instance-level features and object motions. IA-MOT includes three steps: embedding feature extraction, online object tracking with STR, and object re-identification with motion consistency. A Mask R-CNN (He et al., 2017) with an additional embedding head and spatial attention first generate discriminating features. The following MOT stage consists of online Hungarian assignment, short-term retrieve module and ReID.

Authors of GMPHD-SAF (Song et al., 2020) proposed a highly practical online MOTS method which is based on the GMPHD filter and consists of Hierarchical data association (HDA), mask merging, and simple affinity fusion (SAF). Tracking by Segmentation method (Liu et al., 2020) is an offline multistage approach using object embeddings and optical flow. The first stage is the proposal generation; the instance segmentation network will process every frame in each video and generate proposals. Then, short tracklets will be generated by connecting corresponding proposals between two consecutive frames. After that short tracklets which belong to the same object are merged using ReID embeddings as visual similarity cues at the final stage.

Siamese Net methodology is typically used for comparing similar instances in different type sets. In (Lee et al, 2020), authors proposed a Siamese Random Forest (RF) framework that combines an accurate RF with a Siamese structure with a high-speed learning and classification. They applied global averaging pooling (GAP) to feature maps obtained from different layers of darknet53, the backbone network of YOLOv3, to reduce the computation time for feature extraction. TransTrack (Sun et al, 2020) takes advantage of query-key mechanism and introduces a set of learned object queries into the pipeline to enable detecting new-coming objects using brand new architecture based on Transformer. They used idea from Siamese Nets in order to match detection and tracking bounding boxes.

It should be noted that modern image-based approaches to object detection and tracking demonstrate the need for both increasing the speed of instance segmentation and increasing the reliability of forming and matching object embeddings in complex real road scenarios.

3. TASK STATEMENT

In this paper, we consider instance segmentation and tracking methods on two-dimensional images obtained by the monocular camera of an unmanned vehicle. It is necessary to be able to detect objects on an input frame and index them based on objects from previous frames.

It is expected that the model will form individual feature spaces for tracking a specific object using a 3-channel image. In general, an embedding model must be resistant to changes in the object shape, distance, color characteristics. The described limitations significantly affect on developing of a robust tracker model based on deep neural networks. Method should also work in real time.

The main metrics for results comparison will be: sMOTSA, MOTSA, MOTSP, IDS from the KITTI MOTS competition (Voigtlaender et al., 2019).

This work results can be primarily used to construct object trajectories for subsequent position prediction in order to ensure safety.

So, the main task stages are:

1) search and adapt deep neural network state-of-the-art solutions for instance segmentation;

2) build a high-speed image processing pipeline for multi-class object tracking tasks;

3) develop a software implementation using NVidia CUDA technology.

4. METHODOLOGY

The instance segmentation module, based on spatial representations of objects, performs detection. Next, a 2D point cloud is built from the obtained masks, which is then transmitted to the neural network input to obtain vector representations (embeddings) that have a characteristic individual proximity for the same objects (Figure 1).

Instance segmentation methods solve one of the most difficult tasks in computer vision. The main problem is the detection of small objects, accurate prediction of the shape, as well as the class. In the paper, we study Spatial Embeddings (Neven et al., 2019) which is a new method without using region proposals, and also was trained with a new loss function. The algorithm has the encoder-decoder type, the second part of which consists of seed and instance branch. The seed branch learns spatial offsets relative to the centers of objects by class. The instance branch is more complex and consists of several submodules. Sigma map teaches the spatial distribution of objects to take this into account when adjusting the scope for a particular instance.

We perform object tracking using the PointTrack network (Xu et al., 2020). The input is the result of segmentation of instances on a specific frame of the video sequence. The key algorithm attribute is the use of features in the form of "point clouds" to obtain a vector representation of an object using a neural network approach (Figure 1).

There was no class division in the original work (Xu et al., 2020). Since our task is to predict the movement object trajectories to ensure the safety of unmanned vehicles, it is necessary to take into account objects structure and behavior belonging different types separately. The prerequisites for this are the class-dependent form and speed of an instance.

Object tracking is based on the Hungarian algorithm. The similarity measure S (cost matrix) used in the proposed approach is described by:

$$S(C_{si}, C_{sj}) = -D(M_i, M_j) + \alpha \cdot U(C_{si}, C_{sj}),$$

where M_i – embedding of i-th object, D – Euclidian distance between embeddings i-th and j-th, U – intersection-over-union, C_{si} - segmentation mask of i-th object, α - significance coefficient.

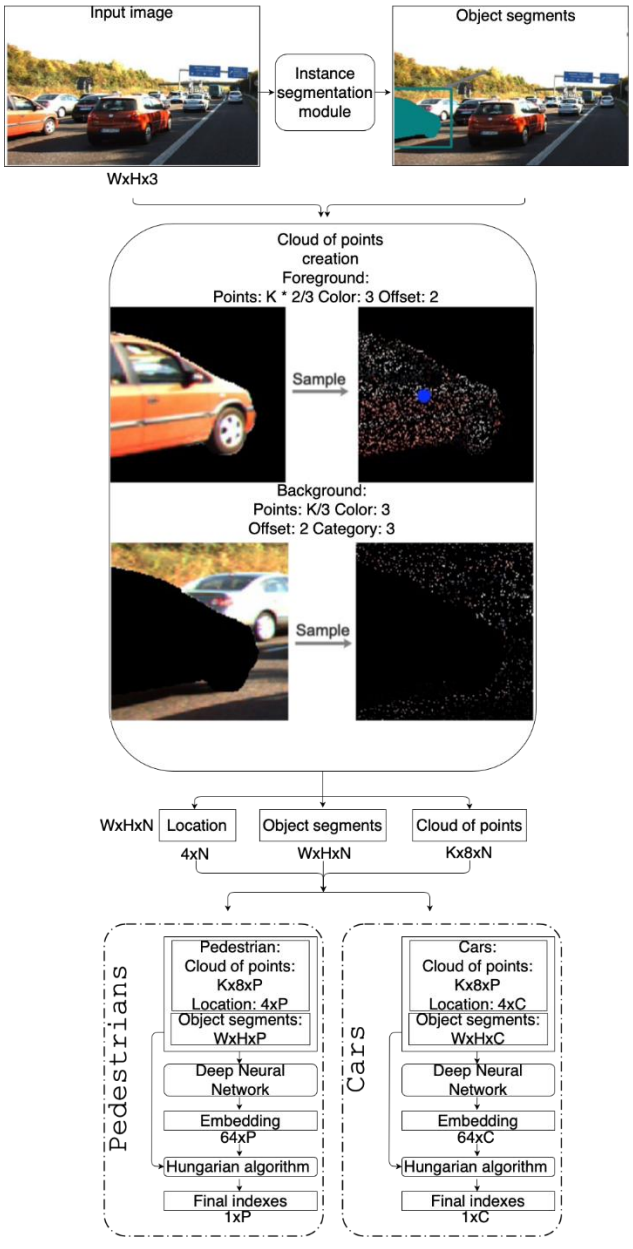

Figure 1. Proposed multi-object tracking and segmentation approach.

After carefully studying the original method results, we spatially limited an object movement between adjacent frames. First, it will help to avoid the problem of re-indexing on distant frames. Second, for similar vector representations over Euclidean distance, we should rely on the actual object location on the frame by storing the gravity center R of the C_{si} object segment:

$$R(C_{si}) = \frac{\sum_{r_j \in C_{si}} r_j}{N},$$

where r_j – the coordinate of a single pixel in the segment.

Despite the minor modifications made to the point cloud-based indexer, an important change in the pipeline operation is the replacement of the segmentation module with a faster one. We use SOLOv2 (Wang et al., 2020) as modern state-of-the-art instance segmentation model. It is customizable, based on the required speed accuracy, you can choose a suitable model.

Next step is training details. The main datasets for model training were KINS (Qi et al., 2019) for segmentation and KITTI MOTS (Voigtlaender et al., 2019) for object tracking. Given the transport specifics of the problem, van, track, and car were selected from original dataset and combined as a single class – car; the pedestrian, cyclist categories were merged into one pedestrian class. Each PointTrack network (Xu et al., 2020) was trained separately for post-processing both pedestrians and cars. The training was carried out in accordance with the original article before the loss function reaching a plateau.

5. EXPERIMENTAL RESULTS

Implementation details. Experiments were performed on a workstation with CPU Intel Xeon 6154 32×3GHz, GPU NVidia TeslaV100 32GB. Deep neural network SOLOv2 was trained on KINS dataset using pretrained weights on COCO dataset. StepLR was chosen as learning rate scheduler. Initial value of learning rate was 0.05. PointTrack methods were trained separately by class on KITTI MOTS dataset without using pretrained weights. The other characteristics of the training methodology are the same. The training data was prepared in accordance with the Figure 1.

Quality improvement. Note the impact of segmentation quality on the tracking results. Recall that object tracking is directly based on the points that are contained inside and outside the object mask. Therefore, if the segmentation result has poor quality, the object features contain the environment features. Thus, the resulting point cloud will be non-uniform, and an object representation vector will significantly differ from the embedding of the same object in the previous frame, and the ID will be assigned incorrectly.

Also, we note the impact of modifications within the ID. The sequence 0002 in the original tracker version clearly shows that after a certain number of frames, the object IDs are repeated. The modification mentioned in the Methodology section solves this problem.

The methods were compared between the modifications of the segmentation module and the center of mass estimation. It has been proven that by keeping the original model architecture and applying the center of mass estimation, we increase the quality of tracking on the same data (Table 1).

Metrics	SOLOv2 ResNet34+ PointTrack (original)	Spatial Embeddings+ PointTrack (original)		SOLOv2 ResNet34+ PointTrack (modified)		Spatial Embeddings +PointTrack (modified)
class	Car	Car	Ped	Car	Ped	Car
sMOTSA	63.10	83.87	61.50	63.54	28.16	**85.09**
MOTSA	77.82	93.31	76.50	77.82	44.01	**94.53**
MOTSP	82.88	**90.26**	81.00	83.13	70.88	**90.26**
Recall	86.01	**96.86**	79.00	84.63	54.44	**96.86**
Precision	92.14	**98.32**	97.90	93.69	87.18	**98.32**
IDS	69	152	176	89	81	**54**

Table 1. Quality comparison between methods through classes on KITTI MOTS validation dataset

Performance study. The high-speed tracking algorithm makes mistakes when the segmentation result is of poor quality. It can be noted that, firstly, SOLOv2 (Wang et al., 2020) crops the object mask at the edges of the image, secondly, for small objects the resulting mask changes strongly over time (Figure 4). It causes problems in further cost matrix computations for association. Thirdly, the objects standing next to others may be

merged into one as in the last picture of Figure 2 or grab part of other object (see Figure 3). Such segmentation model disadvantages impose a limitation on the qualitative tracker analysis when indexing objects.

To increase performance, an acceleration was made in the point cloud creation. The point clouds are formed as multidimensional tensors in the pyTorch framework using NVidia CUDA technology. In the original version, the computation was done on the CPU based numpy-array operations (see “Point cloud generation” row in Table 2).

By increasing the speed of the model due to the introduction of SOLOv2 (Table 2), we slightly degrade the quality of the mask and, as a result, the IoU metric, which is the main member in sMOTSA, MOTSP and others (Table 1). Although, IDS remains at the same level due to the object mass center estimation.

Step of segmentation and tracking algorithm	SOLOv2 ResNet34 + PointTrack (modified)	Spatial Embeddings + PointTrack (original)
Instance segmentation, sec	**0.068 ± 0.020**	0.115 ± 0.009
Object clustering, sec	-	0.100 ± 0.040
Point cloud generation, sec	**0.011 ± 0.010**	0.019 ± 0.010
Neural network for object embedding, sec	**0.005± 0.003**	**0.005± 0.003**
Hungarian algorithm, sec	**0.008 ± 0.007**	**0.008 ± 0.007**
Overall latency, sec	**0.092**	0.247

Table 2. Performance comparison between methods

Figure 2. Multiclass results of modified SOLOv2 ResNet34 + PointTrack on Track 0007 from KITTI MOTS

Figure 3. Blobs on instance segmentation mask

(a) SOLOv2 ResNet34 + PointTrack (modified)

(b) Spatial Embeddings + PointTrack (original)

Figure 4. Tracking results comparison between methods on 0002 track

Following Table 2, the modifications made really speed up the execution time of the presented object tracker based on SOLOv2. Firstly, SOLOv2 has not got an additional time-consuming object clustering, secondly, the method works faster. Thus, the time spent on the execution of multiple objects tracking and segmentation decreased by almost 3 times.

6. CONCLUSIONS

We made improvement of the approach to segmentation and object tracking with multi-stage modular architecture that allows us to get results in real time. A fast and multiclass model architecture based on the SOLOv2 segmentation module was proposed. SOLOv2 provides an average processing time of 68 ms per image, the embedding and tracking module - 24 ms using the NVidia Tesla V100 GPU.
We have demonstrated that the estimation of the mass center for recognized 2D object segment improves the quality of the method.
This indicates that the proposed solution is promising for on-board computer vision systems for both unmanned vehicles and various robotic platforms.
In the future, the proposed tracking pipeline is planned to include faster and more accurate segmentation methods for increasing the object tracking quality. It is prospective to use for this image depth and time-spatial constrains of video sequences.

ACKNOWLEDGEMENTS

This work was supported in part of theoretical investigation and methodology (sections 1-3, 5) by the Government of the Russian Federation under Agreement No 075-02-2019-967 and in part of experimental evaluation (section 4) by the Integrant LLC.

REFERENCES

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B., 2016: Simple online and realtime tracking. In 2016 *IEEE International Conference on Image Processing (ICIP)*, 3464-3468.

Bochinski, E., Eiselein, V., Sikora, T., 2017: High-speed tracking-by-detection without using image information. *14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 1-6.

Cai, J., Wang, Y., Zhang, H., Hsu, H. M., Ma, C., Hwang, J. N., 2020: IA-MOT: Instance-Aware Multi-Object Tracking with Motion Consistency. *arXiv preprint arXiv:2006.13458*.

Chu, P., Ling, H., 2019: Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. *In Proceedings of the IEEE International Conference on Computer Vision*, 6172-6181.

Danelljan, M., Bhat, G., Khan, F. S., Felsberg, M., 2019: Atom: Accurate tracking by overlap maximization. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4660-4669).

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017: Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961-2969).

Hu, A., Kendall, A., Cipolla, R., 2019: Learning a Spatio-Temporal Embedding for Video Instance Segmentation. *arXiv preprint arXiv:1912.08969.*

Kim, A., Ošep, A., Leal-Taixé, L. EagerMOT: Real-time 3D Multi-Object Tracking and Segmentation via Sensor Fusion. *5th BMTT MOTChallenge Workshop*.

Kuhn, H. W., 1955: The Hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2), 83-97.

Lee, J., Kim, S., Ko, B. C., 2020: Fast multiple object tracking using Siamese random forest without online tracker updating. *In Proceedings of the IEEE International Conference on Computer Vision* BMTT Workshop (CVPRW), 1-4.

Liu, Y., Wang, L., Zhao, Y., Shen, H., Ye, J., 2020: Tracking by Segmentation: Person-ReID and Optical Flow Based Offline Tracker for the MOTSChallenge 2020. *5th BMTT MOTChallenge Workshop*.

Luiten, J., Fischer, T., Leibe, B., 2020: Track to reconstruct and reconstruct to track. *IEEE Robotics and Automation Letters*, 5(2), 1803-1810.

Luiten, J., Voigtlaender, P., Leibe, B., 2018: Premvos: Proposal-generation, refinement and merging for video object segmentation. *In Asian Conference on Computer Vision*, 565-580). Springer, Cham.

Neven, D., Brabandere, B. D., Proesmans, M., Gool, L. V., 2019: Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8837-8845.

Payer, C., Štern, D., Neff, T., Bischof, H., Urschler, M., 2018: Instance segmentation and tracking with cosine embeddings and recurrent hourglass networks. *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, 3-11.

Porzi, L., Hofinger, M., Ruiz, I., Serrat, J., Bulo, S. R., Kontschieder, P., 2020: Learning Multi-Object Tracking and Segmentation from Automatic Annotations. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6846-6855.

Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J., 2019: Amodal instance segmentation with kins dataset. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3014-3023.

Song, Y. M., Jeon, M., 2020: Online Multi-Object Tracking and Segmentation with GMPHD Filter and Simple Affinity Fusion. *arXiv preprint arXiv:2009.00100*.

Staroverov, A., Yudin, D. A., Belkin, I., Adeshkin, V., Solomentsev, Y. K., Panov, A. I., 2020: Real-Time Object Navigation With Deep Neural Networks and Hierarchical Reinforcement Learning. *IEEE Access*, 8, 195608-195621.

Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Luo, P., 2020: TransTrack: Multiple-Object Tracking with Transformer. a*rXiv preprint arXiv:2012.15460.*

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., Leibe, B., 2019: MOTS: Multi-object tracking and segmentation. *In Proceedings of the IEEE conference on computer vision and pattern recognition*, 7942-7951.

Wang, X., Zhang, R., Kong, T., Li, L., Shen, C., 2020: SOLOv2: Dynamic and fast instance segmentation. *Advances in Neural Information Processing Systems, 33.*

Xu, Z., Zhang, W., Tan, X., Yang, W., Huang, H., Wen, S., Huang, L., 2020: Segment as points for efficient online multi-

object tracking and segmentation. *In European Conference on Computer Vision*, 264-281.

Yang, F., Chang, X., Dang, C., Zheng, Z., Sakti, S., Nakamura, S., Wu, Y., 2020: ReMOTS: Self-supervised refining multi-object tracking and segmentation. a*rXiv preprint, arXiv: 2007.03200.*

Yuan, Y., Su, X., Zhang, W., Wang, T., Shi, W., Xu, Z., Ding, E., 2020: Re-Identification and Tracklet-Plane Matching for Multi-Object Tracking and Segmentation. *5th BMTT MOTChallenge Workshop*.

Yudin, D., Sotnikov, A., Krishtopik, A., 2019: Detection of Big Animals on Images with Road Scenes using Deep Learning. *In 2019 International Conference on Artificial Intelligence: Applications and Innovations (IC-AIAI)*, 100-103.

Zhang, H., Wang, Y., Cai, J., Hsu, H. M., Ji, H., Hwang, J. N., 2019: LIFTS: Lidar and Monocular Image Fusion for Multi-Object Tracking and Segmentation. *5th BMTT MOTChallenge Workshop*.