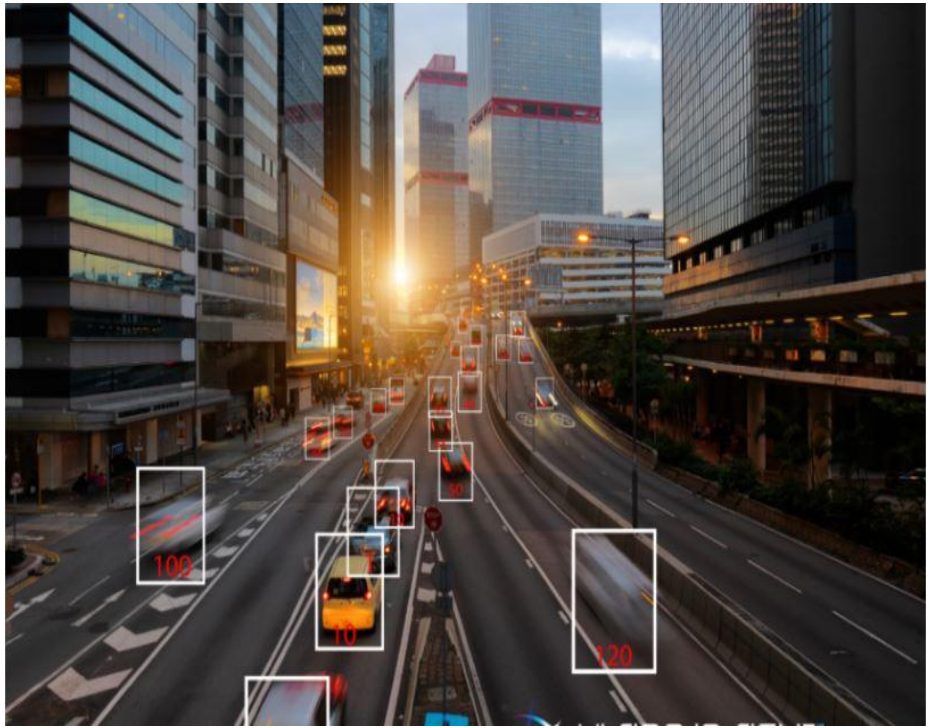
Multi-Object Tracking (MOT)

- MOT, **is** an experimental technique used to study how our visual system tracks **multiple** moving **objects.**

- An essential task in computer vision field.



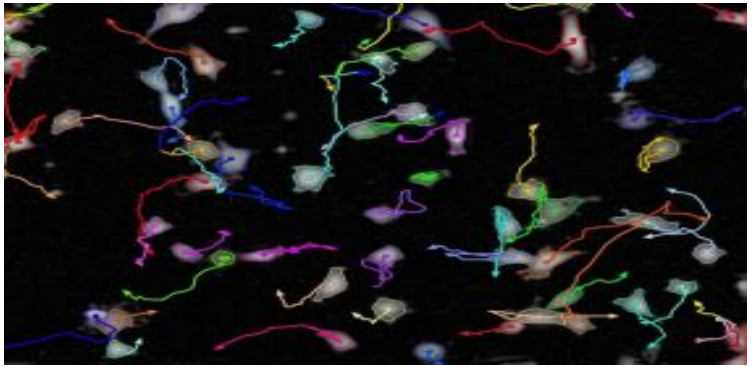
Applications

Autonomous Driving

Sport analytics

Video Surveillance

Medical Research



Main Challenges

Similar Objects

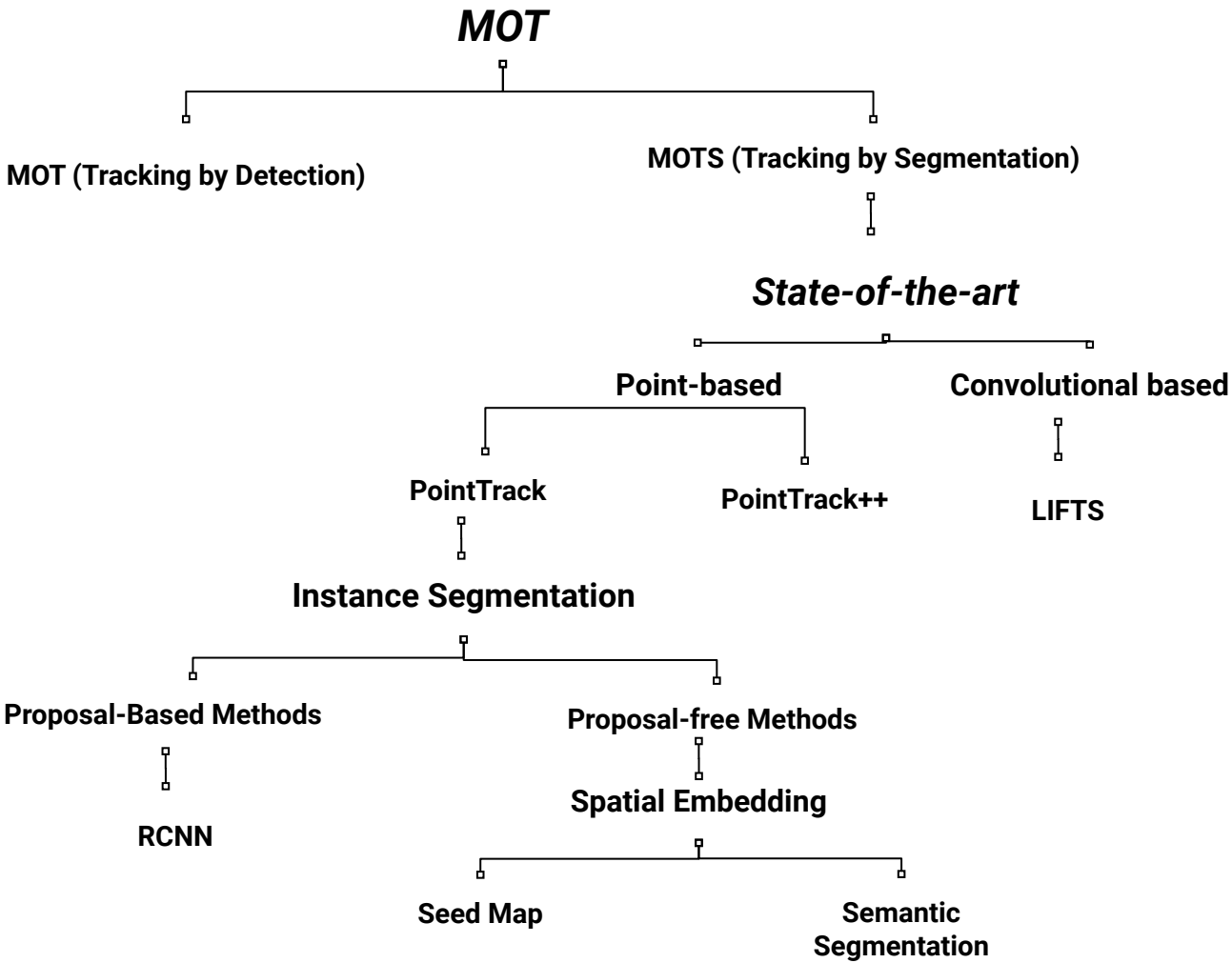
Occlusions

Appearance and pose

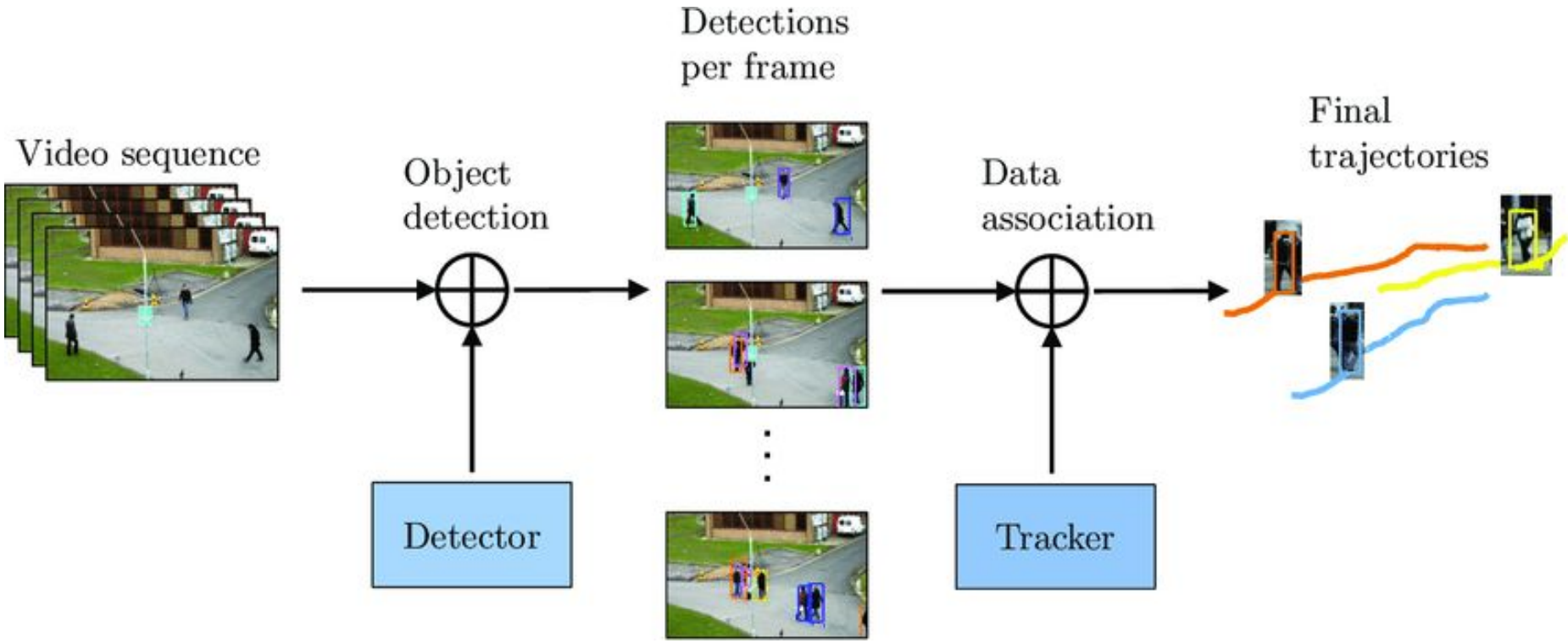
Scale changes

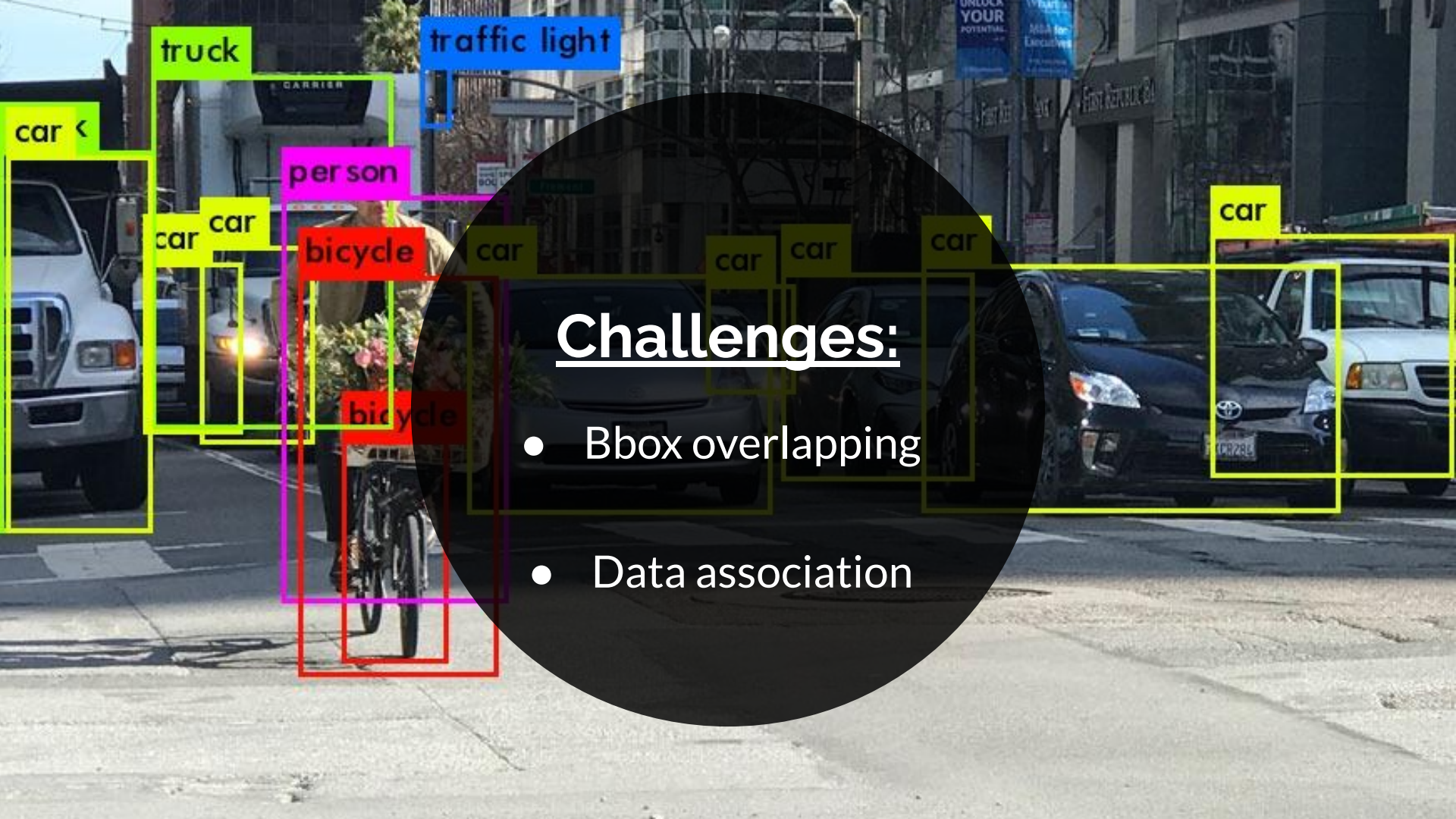
Objective

- **Robust long term tracking while:**
  - Crowded Scenes
  - Shape complexity
  - Edge ambiguity
  - Occlusion


MOT
MOT (Tracking by Detection)
MOTS (Tracking by Segmentation)
State-of-the-art
Point-based
Convolutional based
PointTrack
PointTrack++
LIFTS
Instance Segmentation
Proposal-Based Methods
Proposal-free Methods
RCNN
Spatial Embedding
Seed Map
Semantic Segmentation

Tracking By Detection



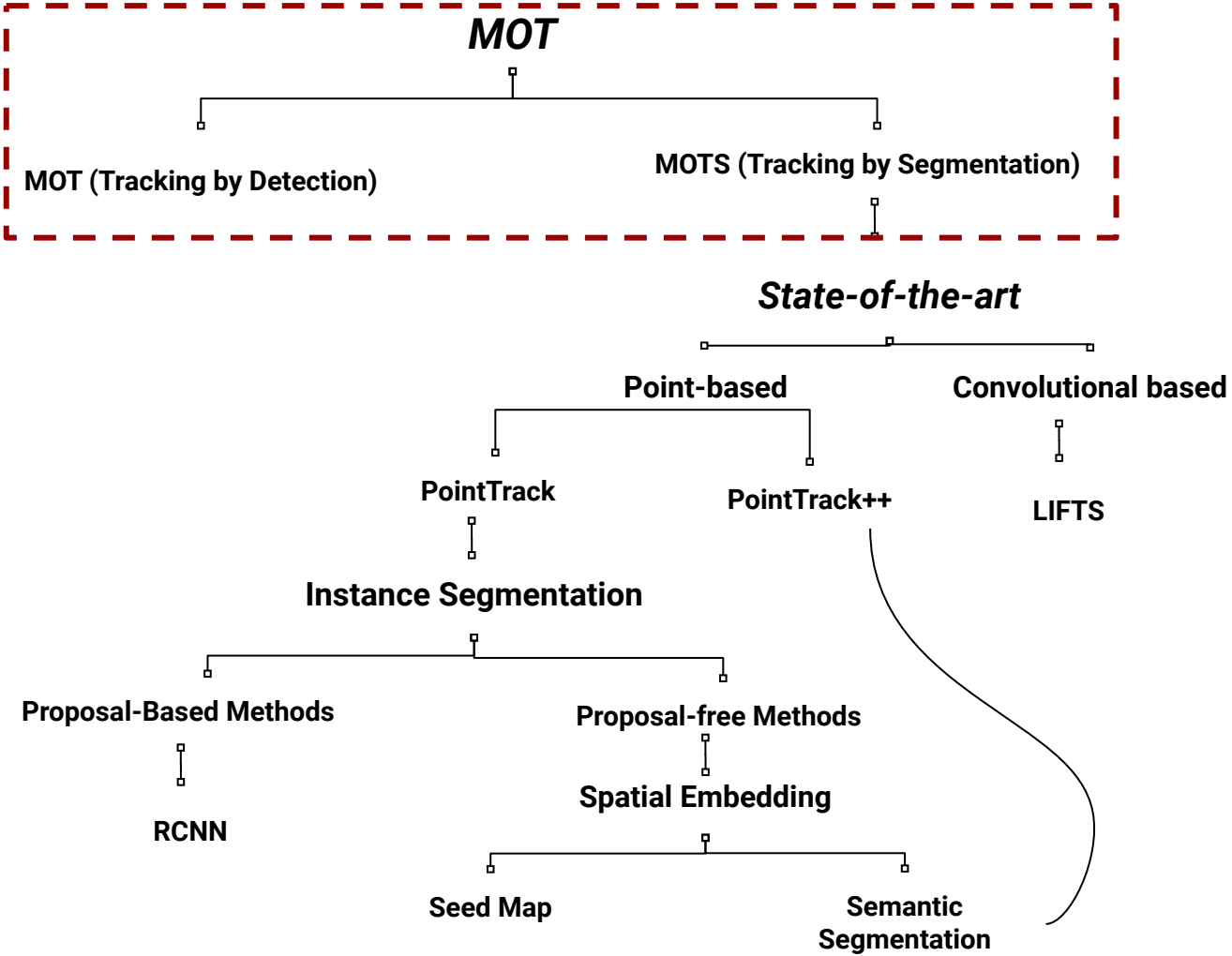


Challenges:

- Bbox overlapping
- Data association

Tracking By Segmentation

- instance masks.
- pixel-level analysis.
- Similarity measurements.


MOT
MOT (Tracking by Detection)
MOTS (Tracking by Segmentation)
State-of-the-art
Point-based
Convolutional based
PointTrack
PointTrack++
LIFTS
Instance Segmentation
Proposal-Based Methods
Proposal-free Methods
RCNN
Spatial Embedding
Seed Map
Semantic Segmentation

LIFTS

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 informa-

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.

LIFTS

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

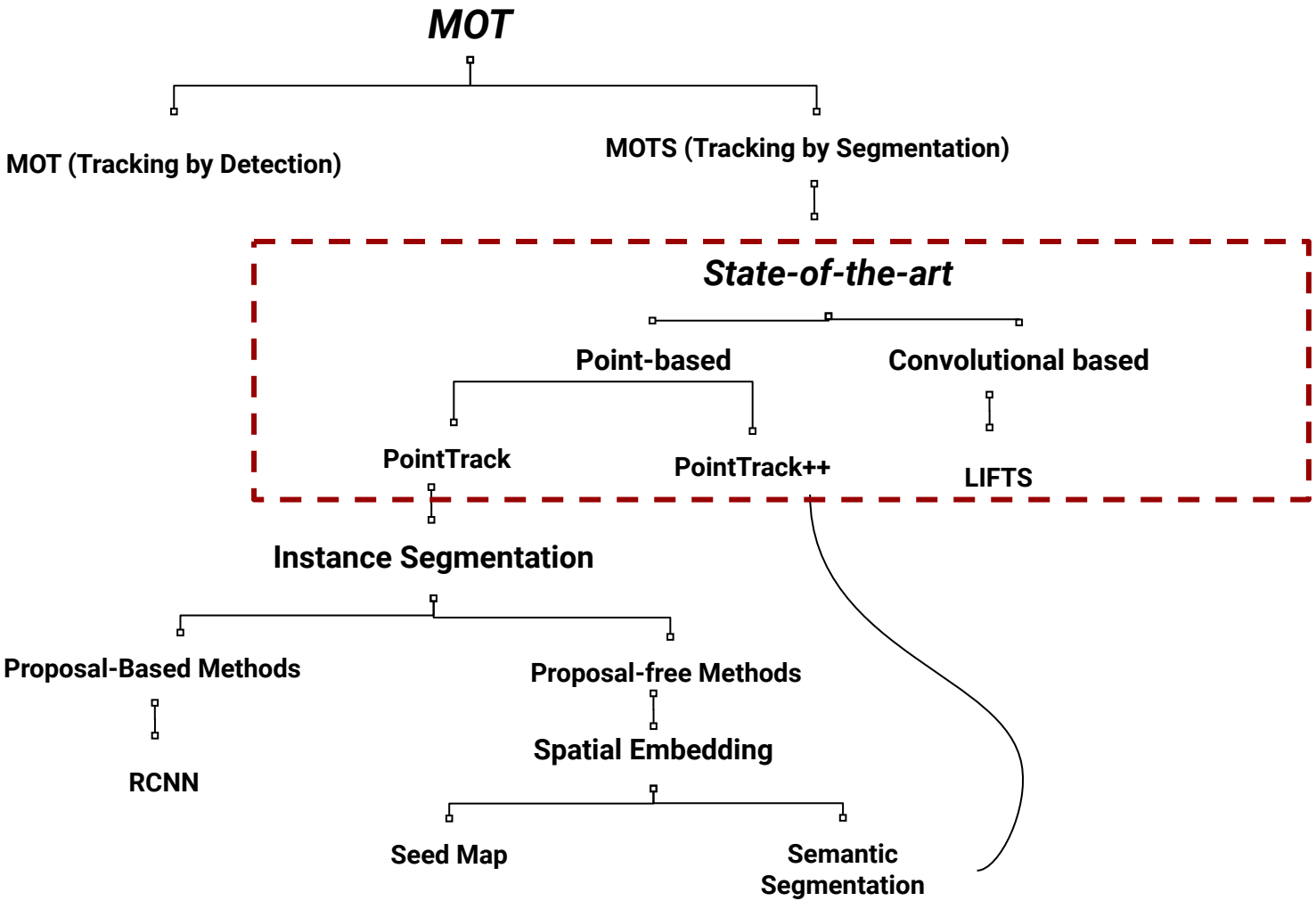
Challenges in Tracking By Segmentation?

Extracting instance feature embeddings from segments

- State-of-the-art problem.
- Convolution in bbox.
- Mixing up features.
- Previous problem

Where are we now?

- Current MOTS methods adopt advanced segmentation networks to extract image feature
- Current MOTS fail to learn discriminative instance embeddings
- Discriminative instance essential for robust instance association
- Limited tracking performances


MOT
MOT (Tracking by Detection)
MOTS (Tracking by Segmentation)
State-of-the-art
Point-based
Convolutional based
PointTrack
PointTrack++
LIFTS
Instance Segmentation
Proposal-Based Methods
Proposal-free Methods
RCNN
Spatial Embedding
Seed Map
Semantic Segmentation

PointTrack

Segment as Points for Efficient Online Multi-Object Tracking and Segmentation

Zhenbo Xu[1,2][0000−0002−8948−1589], Wei Zhang[2], Xiao Tan[2], Wei Yang[1*][0000−0003−0332−2649], Huan Huang[1], Shilei Wen[2], Errui Ding[2], and Liusheng Huang[1]

[1] University of Science and Technology of China
[2] Department of Computer Vision Technology (VIS), Baidu Inc., China
[*] Corresponding Author. E-mail: qubit@ustc.edu.cn

Abstract. Current multi-object tracking and segmentation (MOTS) methods follow the tracking-by-detection paradigm and adopt convolutions for feature extraction. However, as affected by the inherent receptive field, convolution based feature extraction inevitably mixes up the foreground features and the background features, resulting in ambiguities

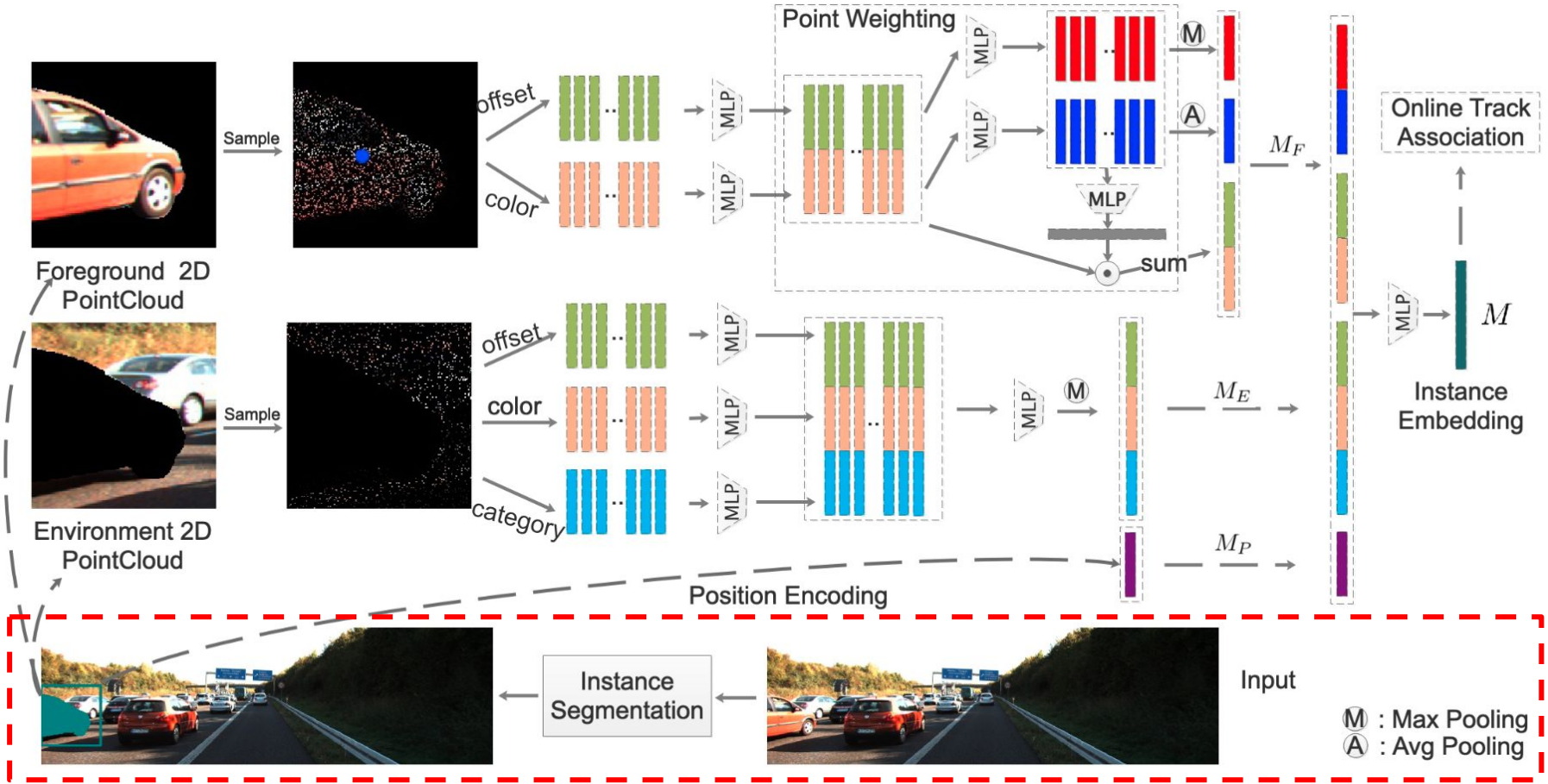
PointTrack

Segment as Points for Efficient Online Multi-Object Tracking and Segmentation

Tracking By Points:

- Foreground pointcloud
- Background pointcloud
- Learn features on them **separately**

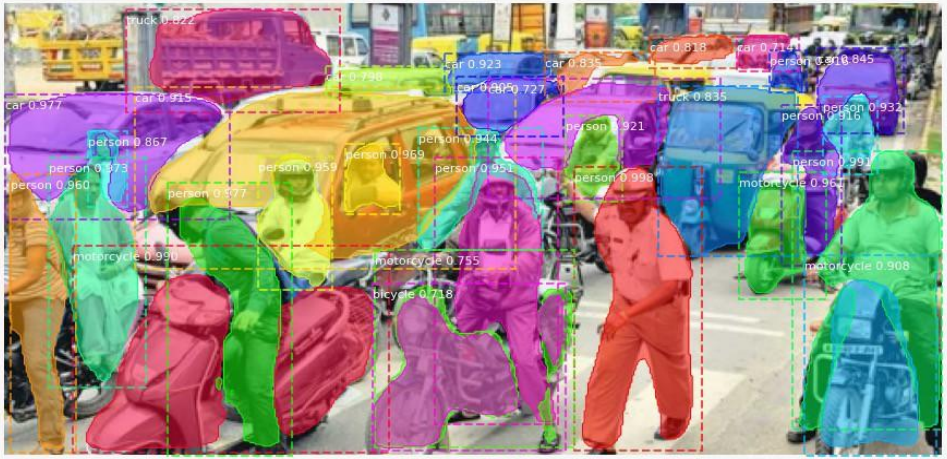
PointTrack: Instance Segmentation

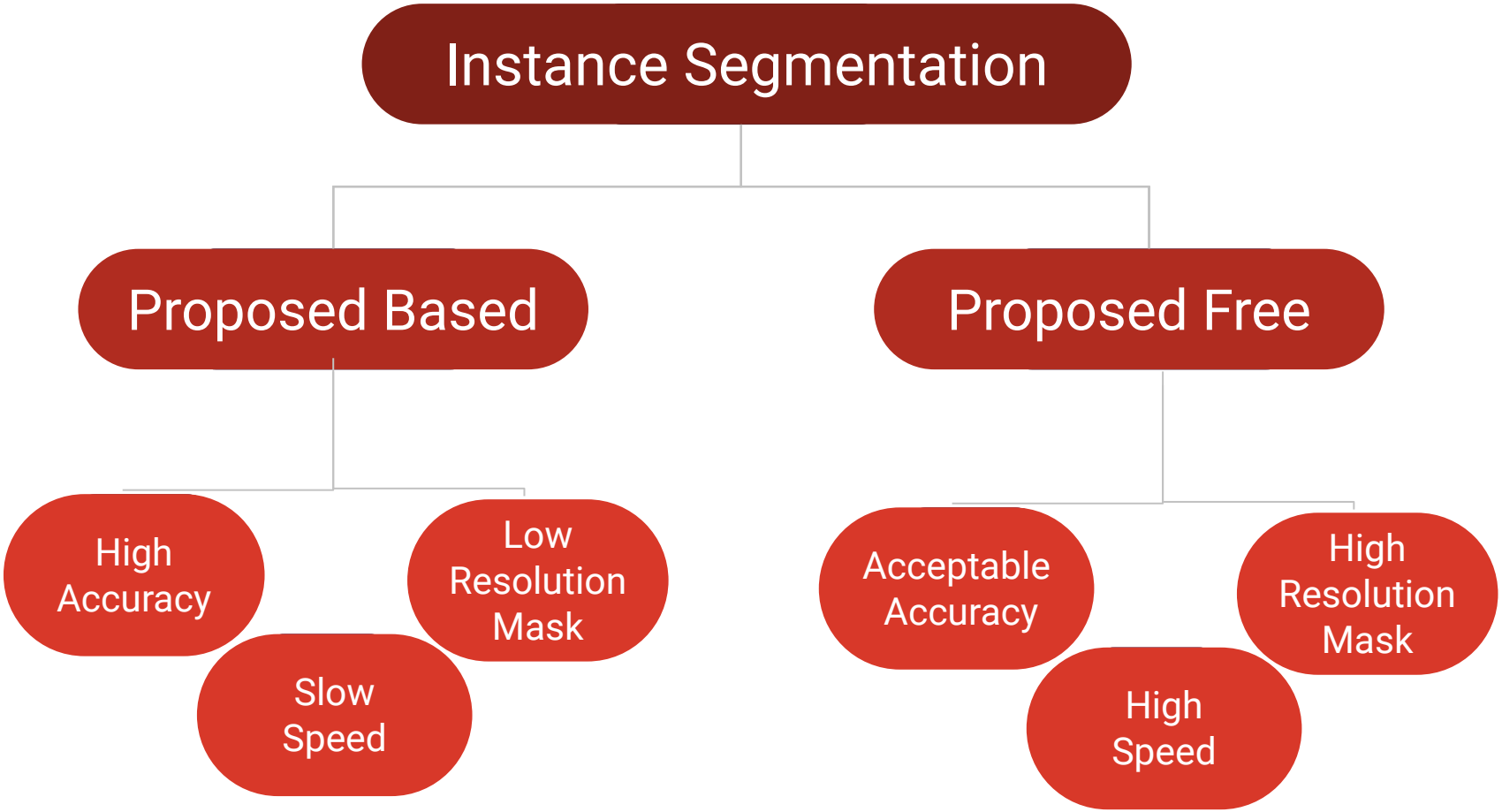


PointTrack: Instance Segmentation

What is be the best for PointTrack?

- Tracking By Points flexibility
- State-of-the-art

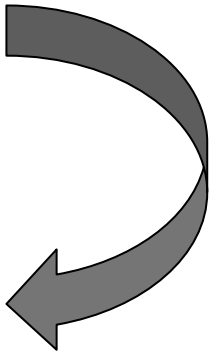



Instance Segmentation
Proposed Based
Proposed Free
High Accuracy
Low Resolution Mask
Slow Speed
Acceptable Accuracy
High Resolution Mask
High Speed

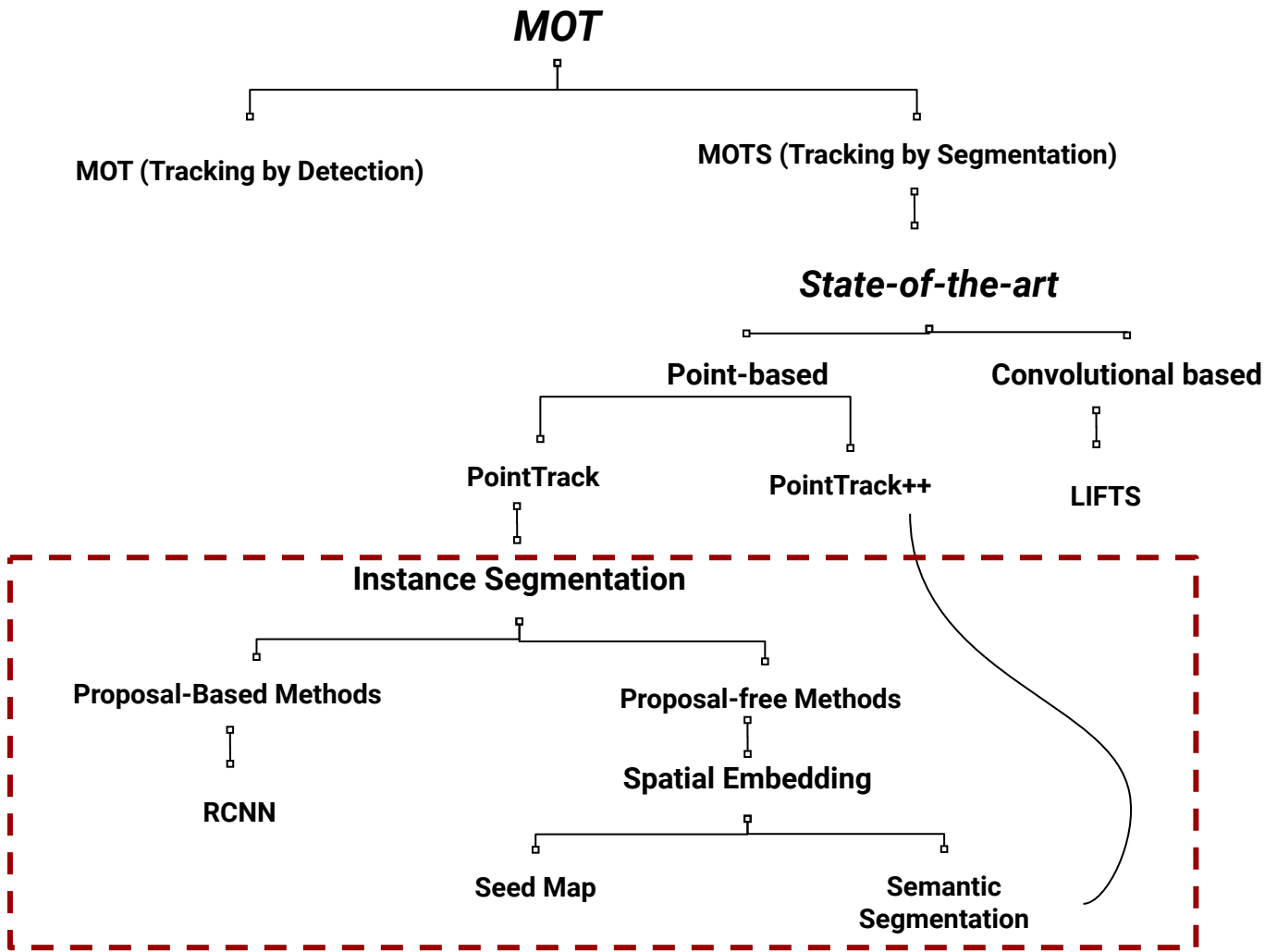
Autonomous Vehicles Challenges

Instance Segmentation requirements:

- High Accuracy (Proposed Based)
- High Resolution Mask
- **Real Time** (Proposed Free)



Trade Off


MOT
MOT (Tracking by Detection)
MOTS (Tracking by Segmentation)
State-of-the-art
Point-based
Convolutional based
PointTrack
PointTrack++
LIFTS
Instance Segmentation
Proposal-Based Methods
Proposal-free Methods
RCNN
Spatial Embedding
Seed Map
Semantic Segmentation

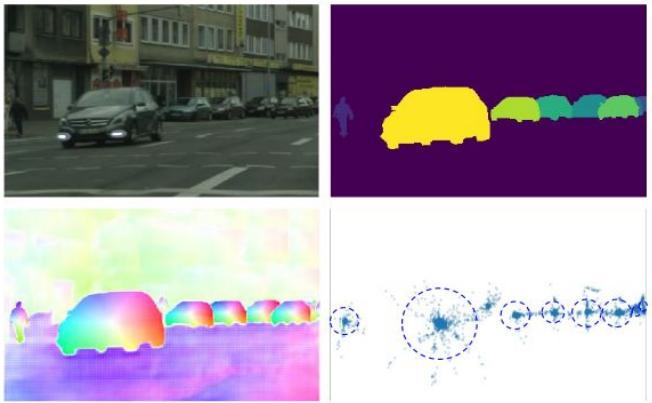
Spatial Embedding

Instance Segmentation by Jointly Optimizing Spatial Embeddings and Clustering Bandwidth

Davy Neven Bert De Brabandere Marc Proesmans Luc Van Gool
Dept. ESAT, Center for Processing Speech and Images
KU Leuven, Belgium
{firstname.lastname}@esat.kuleuven.be

Abstract

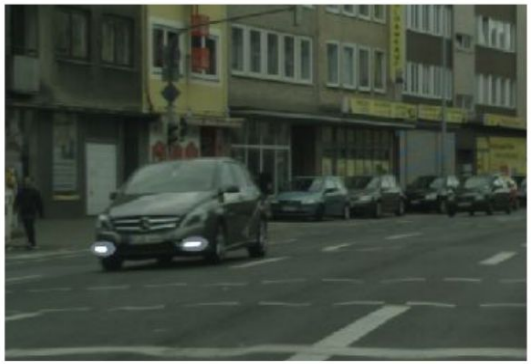
Current state-of-the-art instance segmentation methods are not suited for real-time applications like autonomous driving, which require fast execution times at high accuracy. Although the currently dominant proposal-based methods have high accuracy, they are slow and generate masks at a fixed and low resolution. Proposal-free methods, by contrast, can generate masks at high resolution and are often faster, but fail to reach the same accuracy as the proposal-based methods. In this work we propose a new clustering loss function for proposal-free instance segmentation. The loss function pulls the spatial embeddings of pixels belonging to the same instance together and jointly

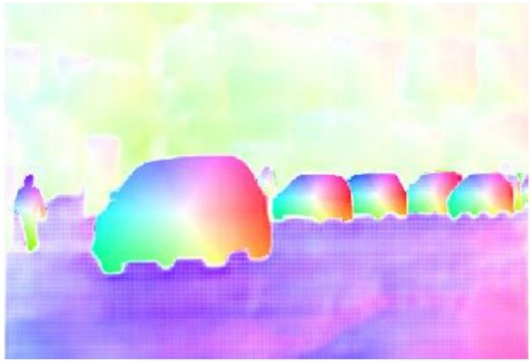


Center Regression

An effective method is learning **offset vectors** for each pixel, **pointing** at the **center** of its object:

$$\mathcal{L}_{regr} = \sum_{i=1}^{n} \|o_i - \hat{o}_i\|$$





Promising given spatial-invariant nature of CNN
However:

1) **Loss (**L2 or L1 norm) does does not directly optimize mask iou - not **end-to-end**
2) Loss **dominated by distance pixels.**

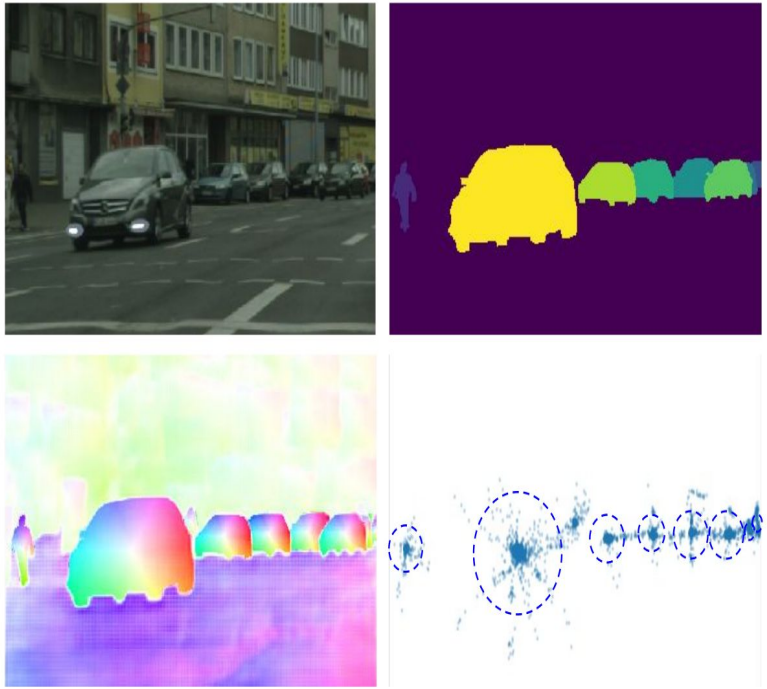
Distance to Probability

1) **Shift** pixel coordinates with predicted **offset vectors**, forming "**Spatial Embedding**".

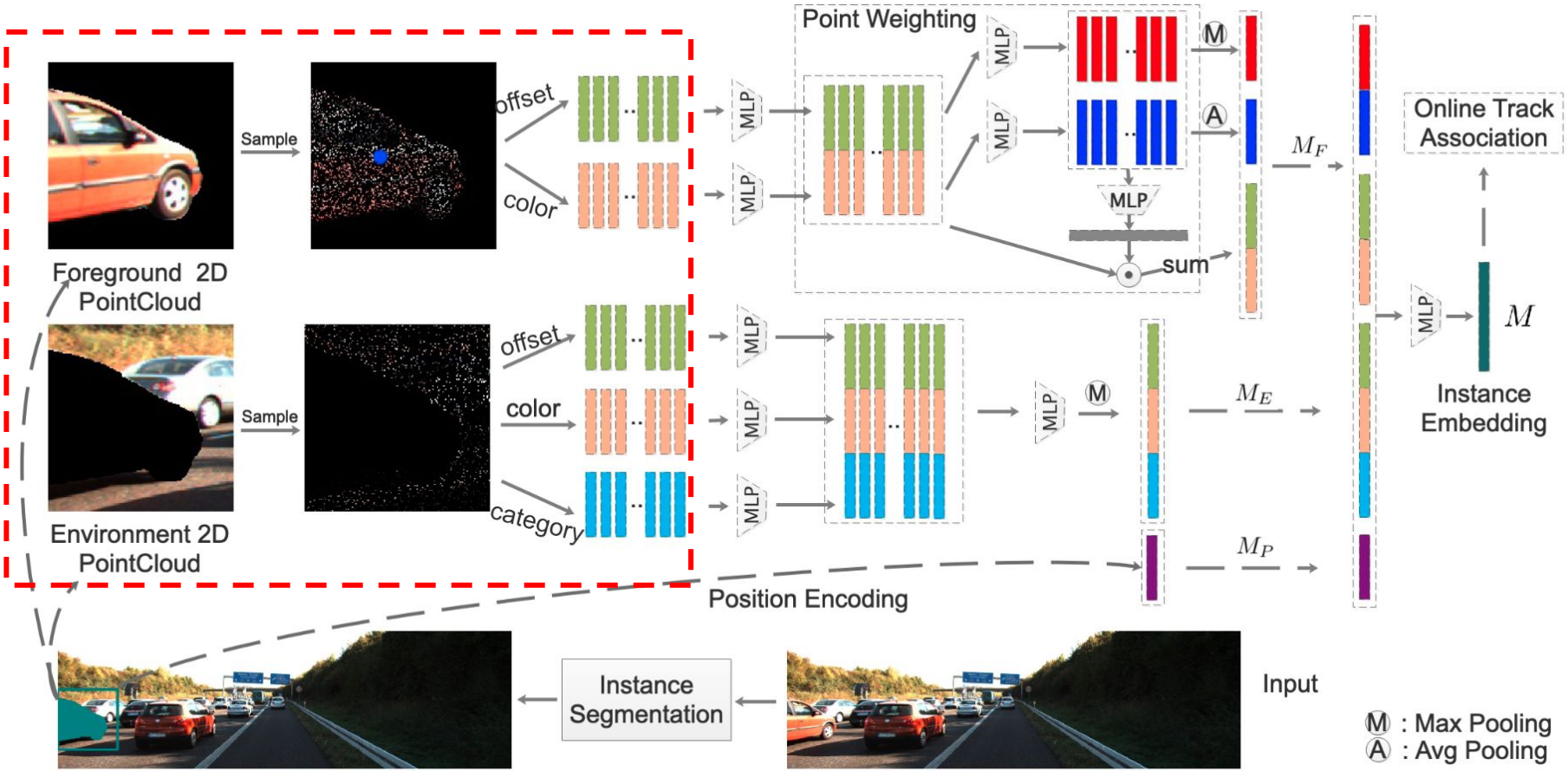
2) Convert embedding-to-center-distance into **probability** of belonging to object:

$$\phi_k(e_i) = \exp\left(-\frac{\|e_i - C_k\|^2}{2\sigma_k^2}\right)$$

3) Optimize this **probability map** with **Hinge**
4) **loss** for maximizing mask iou end-to-end



PointTrack: Extracting Instance Feature Embeddings from Segments



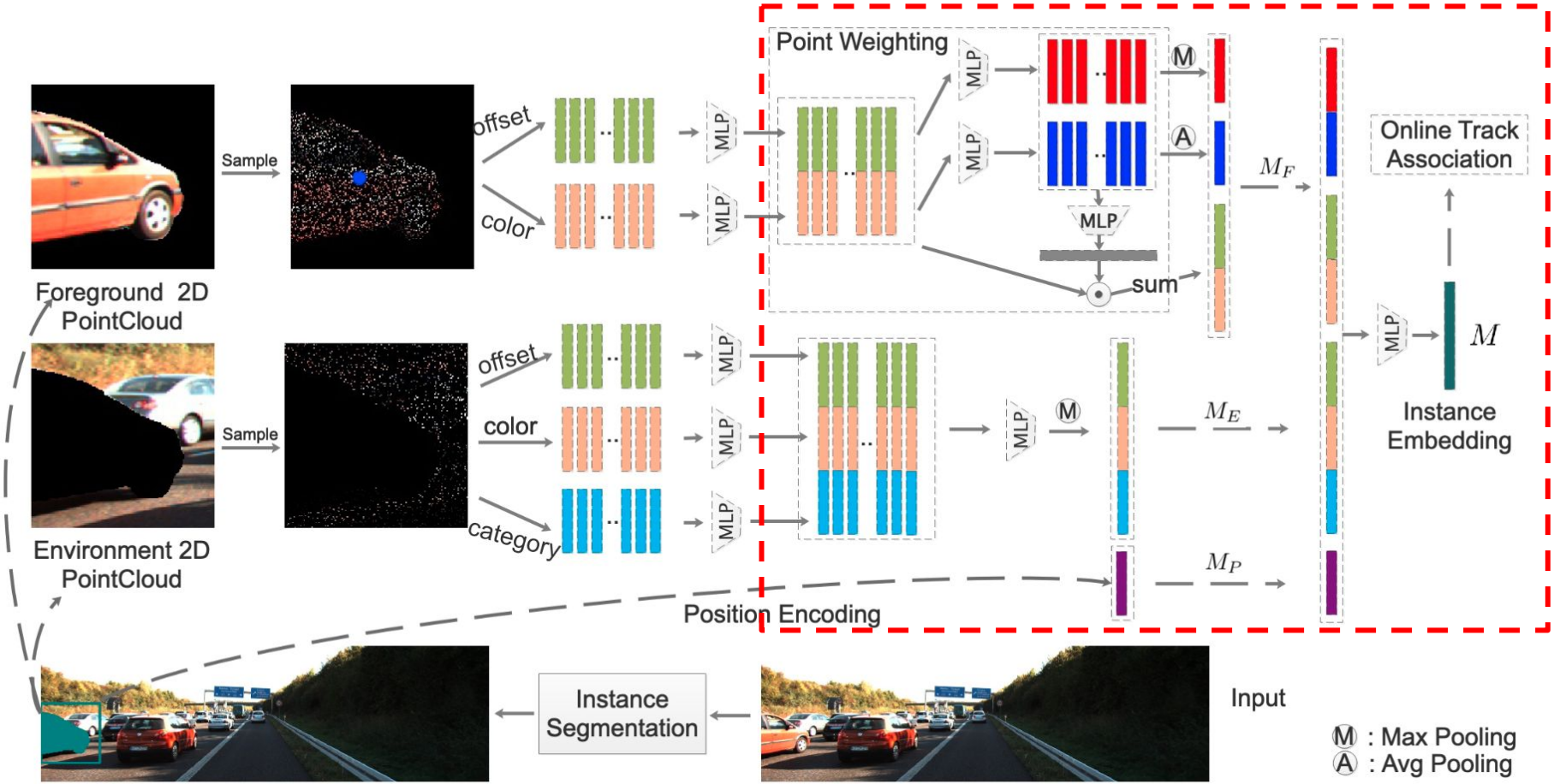
PointTrack: Extracting Instance Feature Embeddings from Segments

Summary:

- **Foreground point-cloud:**
  1. Offset (Scale, Shape)
  2. Color (Appearance)

- **Background point-cloud:**
  1. Offset (Scale, Shape)
  2. Color (Appearance)
  3. Offset + Category (nearby instances aware)

PointTrack: Multi-Layer Perceptron



PointTrack: **Instance Association**

Instance association based on similarities

$$S(C_{s_i}, C_{s_j}) = -D(M_i, M_j) + \alpha * U(C_{s_i}, C_{s_j})$$

D denotes the Euclidean distance.
U represents the mask IOU.
C instance, M Embedding

Intuition:

- If an active track does not update for the recent β frames, we end this track automatically.
- set a similarity threshold γ for instance association and instance association if similarity is greater than γ.
- After instance association, unassigned segments will start new tracks.

PointTrack++

PointTrack++ for Effective Online Multi-Object Tracking and Segmentation

Zhenbo Xu[1], Wei Zhang[2], Xiao Tan[2], Wei Yang[1]*, Xiangbo Su[2], Yuchen Yuan[2],
Hongwu Zhang[2], Shilei Wen[2], Errui Ding[2], Liusheng Huang[1]
[1]University of Science and Technology of China
[2]Department of Computer Vision Technology (VIS), Baidu Inc., China

Abstract

Multiple-object tracking and segmentation (MOTS) is a novel computer vision task that aims to jointly perform multiple object tracking (MOT) and instance segmentation. In this work, we present PointTrack++, an effective online framework for MOTS, which remarkably extends our recently proposed PointTrack framework. To begin with, PointTrack adopts an efficient one-stage framework for instance segmentation, and learns instance embeddings by converting compact image representations to un-ordered 2D point cloud. Compared with PointTrack, our proposed PointTrack++ offers three major improvements. Firstly, in the instance segmentation stage, we adopt a semantic segmentation decoder trained with focal loss to improve the instance selection quality. Secondly, to further boost the segmentation performance, we propose a data augmentation strategy by copy-and-paste instances into training images. Finally, we introduce a better training strategy in the

courages to learn more discriminative embeddings for instance association based on segments rather than bboxes.

Nevertheless, learning instance embeddings from segments have rarely been explored by current MOTS methods. TRCNN [8] extends Mask-RCNN to jointly process consecutive frames using 3D convolutions and adopts ROI Align to extract instance embeddings in bbox proposals. To focus on the segment area, Porzi *et al.* [5] introduce mask pooling rather than ROI Align for instance feature extraction. However, vanilla 2D or 3D convolutions are harmful for learning discriminative instance embeddings due to inherent large receptive fields. Deep convolutional features not only mix up the foreground area and the background area but also mix up the foreground area of the interested instance and its adjacent instances. Therefore, though current MOTS methods adopt advanced segmentation backbones to extract image features, they fail to learn discriminative instance embeddings which are essential for robust instance association, resulting in limited performances.

PointTrack++

PointTrack++ offers three major improvements:

1) instance segmentation, adopt a **semantic segmentation decoder** trained with focal loss to **improve instance association quality**.

2) Further boost the segmentation performance, we propose a data augmentation strategy by **copy-and-paste instances into training images.**

3) Better training strategy in the instance association stage to improve the distinguishability of learned instance embeddings

State-of-the-art Benchmark on KITTI dataset

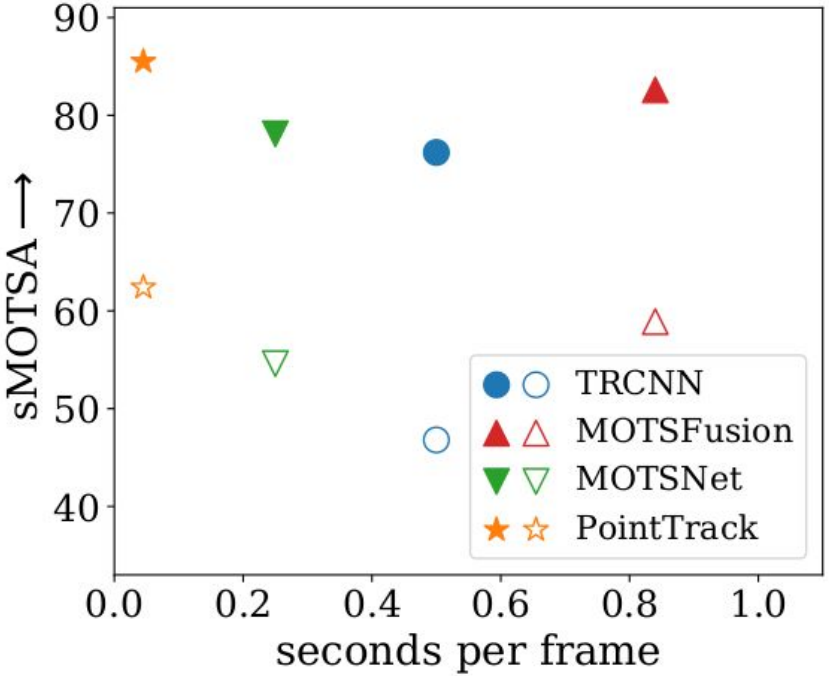
TABLE I
RESULTS ON KITTI DATASET FOR CARS

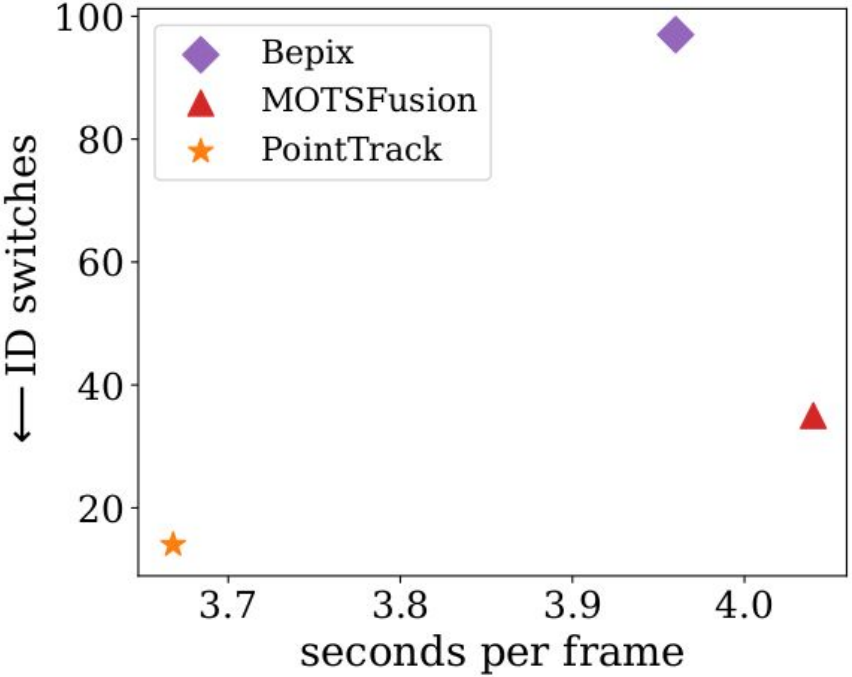
Method	sMOTSA	MOTSA	MOTSP	MODSA	MT	ML	IDS	Frag
LIFTS	79.6	89.6	89	89.9	79.1	2.9	114	532
PointTrack	87.5	90.9	87.1	91.82	90.84	0.6	346	538
PointTrack++	82.8	92.6	89.73	93.35	89.49	1.20	270	579

TABLE II
RESULTS ON KITTI DATASET FOR PEDESTRIANS

Method	sMOTSA	MOTSA	MOTSP	MODSA	MT	ML	IDS	Frag
LIFTS	64.90	80.90	81.00	81.90	61.5	8.90	206	277
PointTrack	61.74	76.5	80.96	77.36	48.9	9.26	176	716
PointTrack++	68.13	83.67	82.22	84.88	66.67	4.81	250	536

State-of-the-art Benchmark on KITTI dataset





Our Evaluation Results:

TABLE VI
MAIN CHALLENGES IN TEST DATASET

Video	Particle Occlusion	Full Occlusimon	Scale Variance	Appearance Variance
Video 2	✓	✓	✓	✓
Video 6	✓	✗	✓	✓
Video 8	✗	✗	✓	✓
Video 13	✓	✗	✓	✗
Video 18	✓	✓	✗	✗

Our Evaluation Results:

TABLE III
EVALUATION MATRIX

Videos	sMOTSA	MOTSA	MOTSP	MOTSAL	MODSA	MODSP	Recall	Prec	F1	FAR	MT	PL
Video2	77.26	91.14	85.22	91.61	91.69	87.16	93.91	97.70	95.77	8.58	100.00	0.00
Video6	89.75	97.77	91.89	97.77	97.77	92.75	98.98	98.88	98.88	2.22	100.00	0.00
Video8	89.06	97.41	91.48	97.57	97.60	91.32	97.98	99.61	98.79	1.03	90.48	9.52
Video13	-9.34	-2.78	92.84	-2.78	-2.78	99.30	91.67	49.25	64.08	10.00	100.00	0.00
Video16	69.80	88.37	79.62	89.00	89.09	79.58	91.13	97.81	94.35	8.13	100.00	0.00

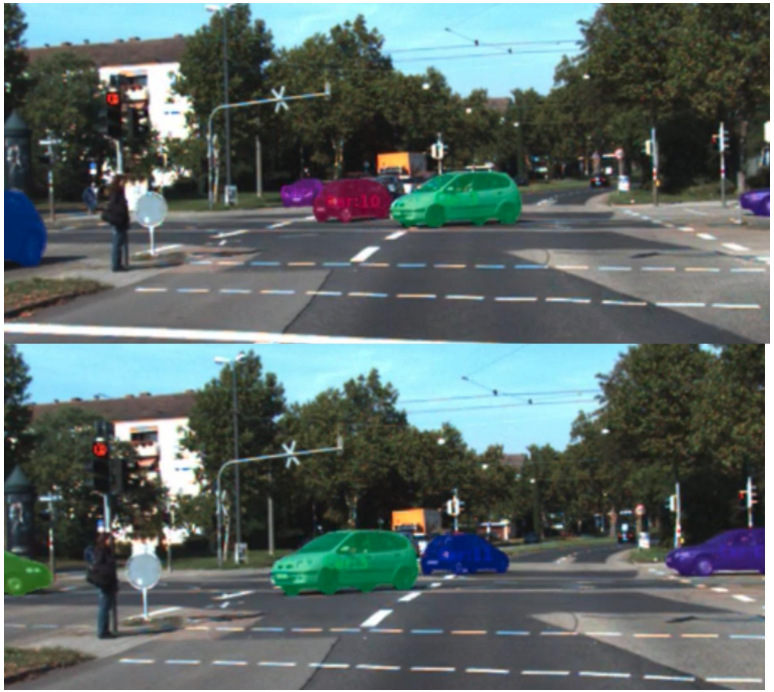
TABLE IV
EVALUATION MATRIX

Videos	ML	TP	FP	FN	IDS	Frag	GT Obj	GT Trk	TR Obj	TR Trk	Ig TR Tck
Video2	0.00	848	20	55	5	26	903	15	1201	11	333
Video6	0.00	1021	4	21	2	10	1042	21	1216	15	191
Video8	0.00	531	6	6	0	3	537	11	756	6	219
Video13	0.00	33	34	3	0	1	36	2	124	5	57
Video16	0.00	760	17	74	6	38	834	4	783	9	6


clideo.com

Remaining Challenges

Id switch after full Occlusion



Tracking Distant car, Detect Truck as car

Conclusion

To conclude,

1) **Tracking by segmentation** shows better performance than tracking by detection which fails during overlapping in the crowded scene.
2) After reviewing the **state of the art papers** for MOTS, we ending up with **PointTrack** approach that propose a **new tracking by point model** which shows high efficiency for instance embedding learning by breaking the image into unordered 2D point-cloud.
3) PointTrack uses APOLLO MOTS dataset for training the model in crowded scene. Moreover, in testing the model produces a good tracking performance in most of the MOT challenges however fails in some situations such as not detecting the far away objects.

Future Work

To overcome the previously mentioned problems:

1) **Tracking small scale objects:**

   a) Extracting a point of interest (POI) in different scale levels using flexible detector descriptor like SURF

   b) Extracting POI instead of random points

   c) Decease PointTrack speed

Future Work

To overcome the previously mentioned problems:

2) Id switch after full occlusion

a) Lidar data could significantly improve the tracking robustness

b) Using prediction model such as kalman filter in full occlusion case

Future Work

To overcome the previously mentioned problems:

3) **Lidar data:**

   a) Depth to the feature vector from the PointTrack

   b) Increase the speed than the state-of-the-art approaches.

   c) However, how to match the lidar points with the same image points?

Reference

1) Zhang, Haotian, et al. "Lifts: Lidar and monocular image fusion for multi-object tracking and segmentation." *BMTT Challenge Workshop, IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

2) Xu, Zhenbo, et al. "Segment as points for efficient online multi-object tracking and segmentation." *European Conference on Computer Vision*. Springer, Cham, 2020.

3) Xu, Zhenbo, et al. "PointTrack++ for Effective Online Multi-Object Tracking and Segmentation." *arXiv preprint arXiv:2007.01549* (2020).

4) Neven, Davy, et al. "Instance segmentation by jointly optimizing spatial embeddings and clustering bandwidth." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

5) pointTrack code reference: https://github.com/detectRecog/PointTrack

Any Questions?

Thank You!