ReMOTS: Refining Multi-Object Tracking and Segmentation (1st Place Solution for MOTS 2020 Challenge 1)

Fan Yang[1,2], *Xin Chang*[1], *Chenyu Dang*[1], *Ziqiang Zheng*[3], *Sakriani Sakti*[1,2], *Satoshi Nakamura*[1,2], *Yang Wu*[4]

[1]Nara Institute of Science and Technology, Japan

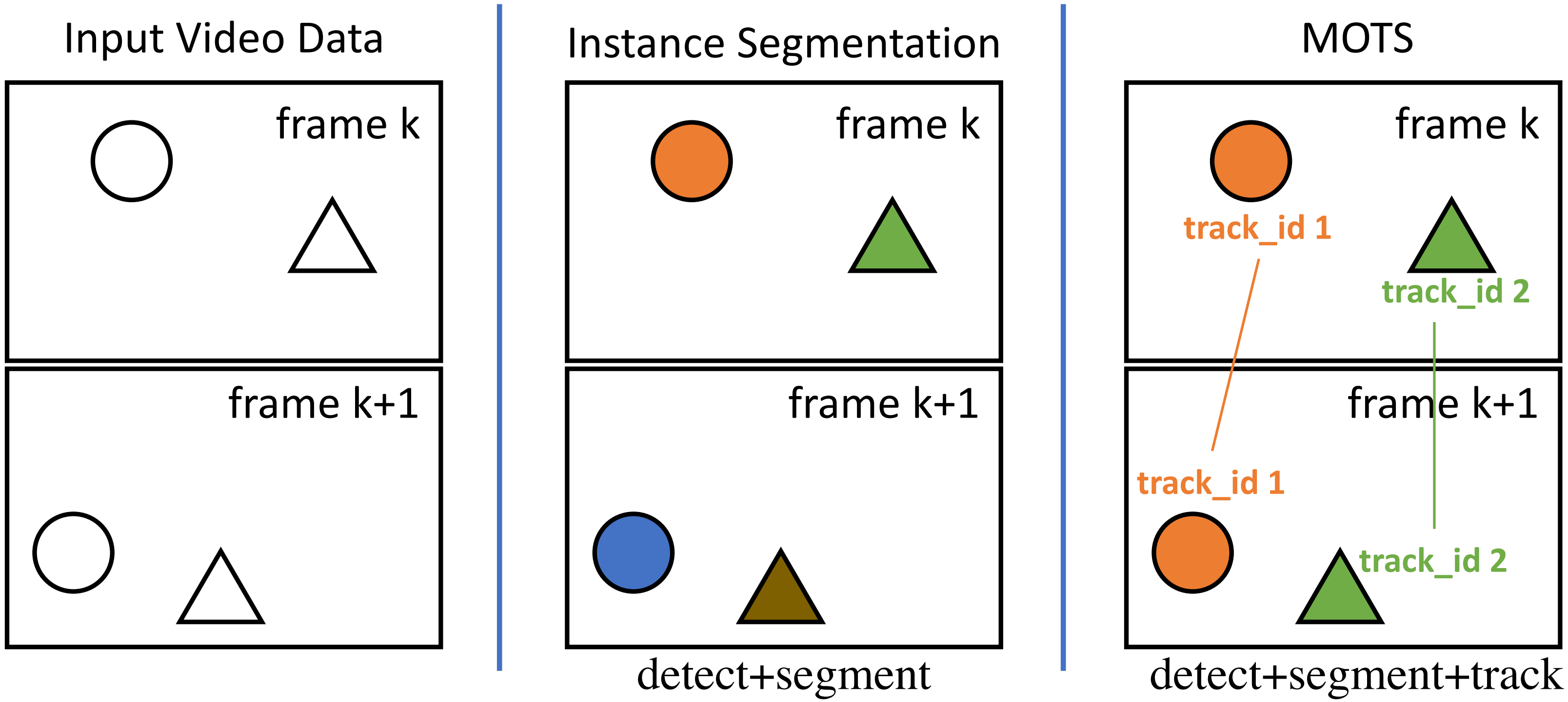
[2]RIKEN Center for Advanced Intelligence Project, Japan

[3]UISEE Technology (Beijing) Co. Ltd., China

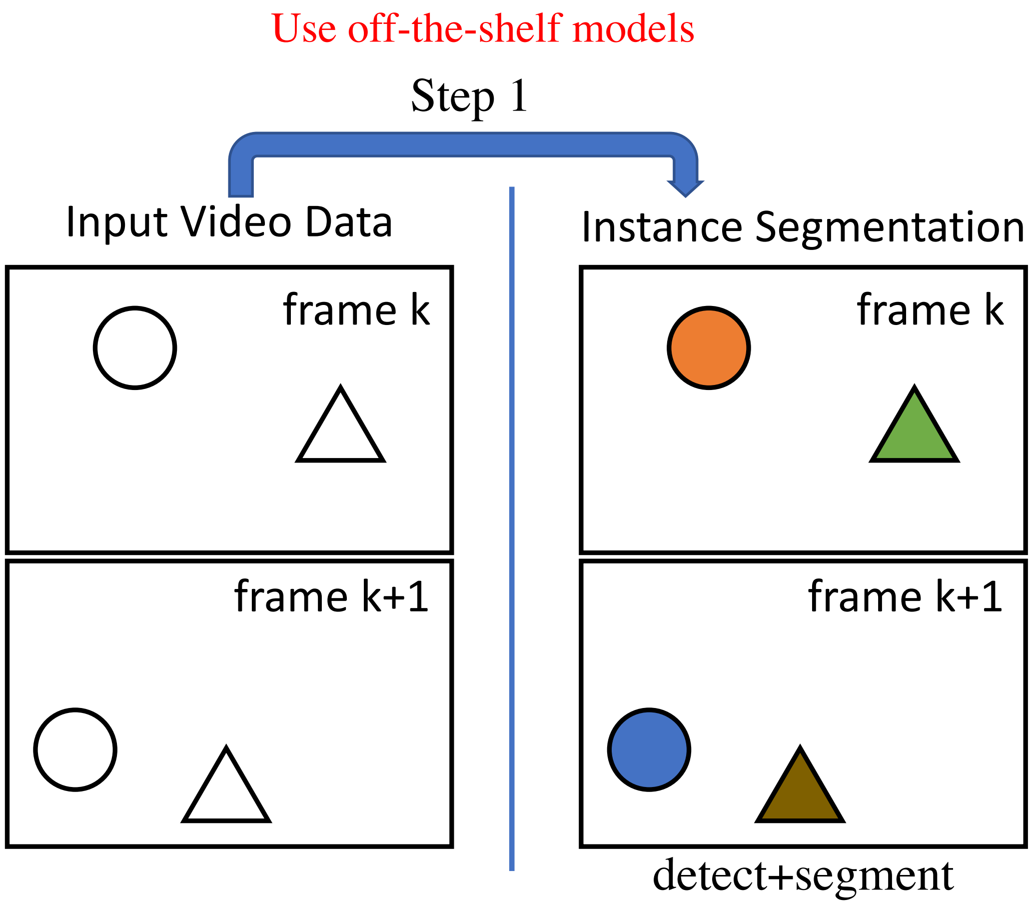
[4]Kyoto University, Japan

Background of Multi-Object Tracking and Segmentation (MOTS)

- Problem: detect, segment, and track multiple objects in videos.
- Input: a video sequence contain that multiple RGB images.
- Output: 2D mask and corresponding track ID at each frame.
- Application: action recognition, automatic driving, and others.



Our solution for MOTS

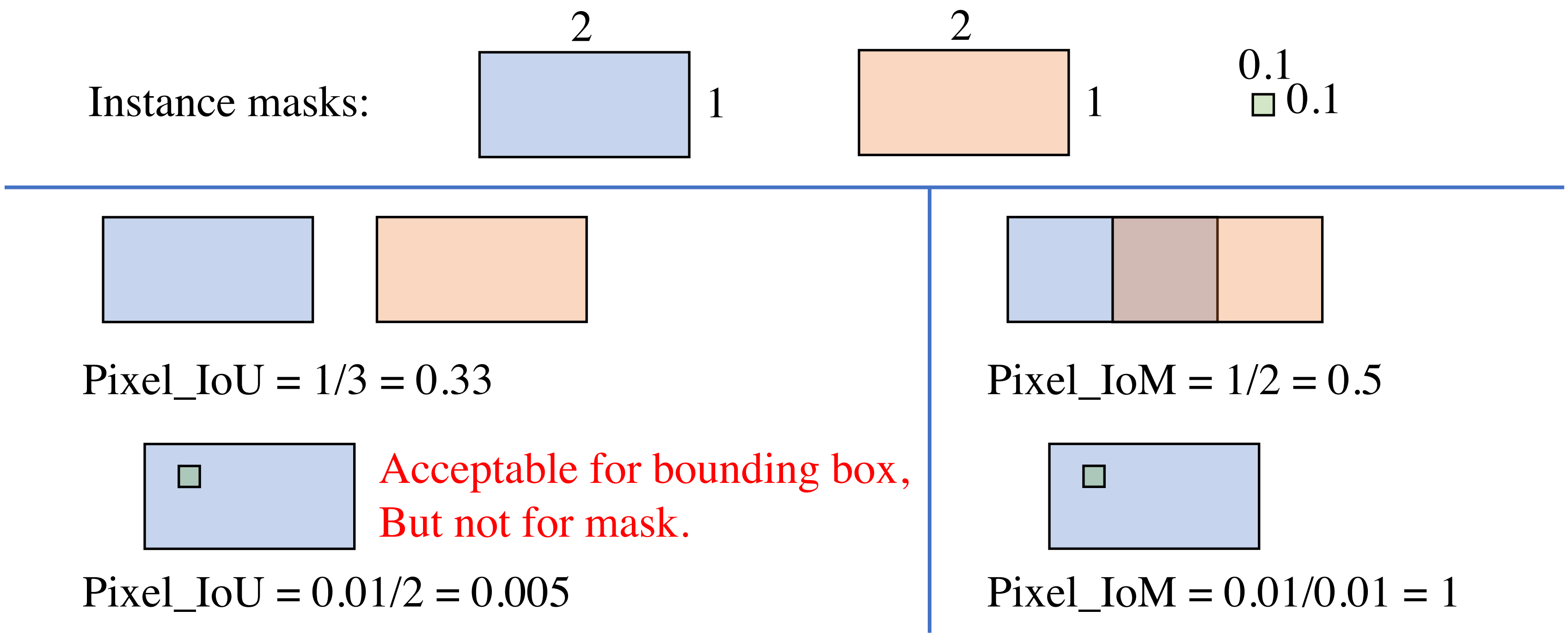


Instance Segmentation

We take off-the-shelf models:
X-101-64x4d-FPN of MMDetection + Mask R-CNN X152 of Detectron 2,
which refers to the public detection and segmentation methods.

But, how to fuse instance masks from different models?

Fusing boxes – using NMS
Fusing masks – may also using NMS – but change IoU to IoM (Intersection over Minimum).

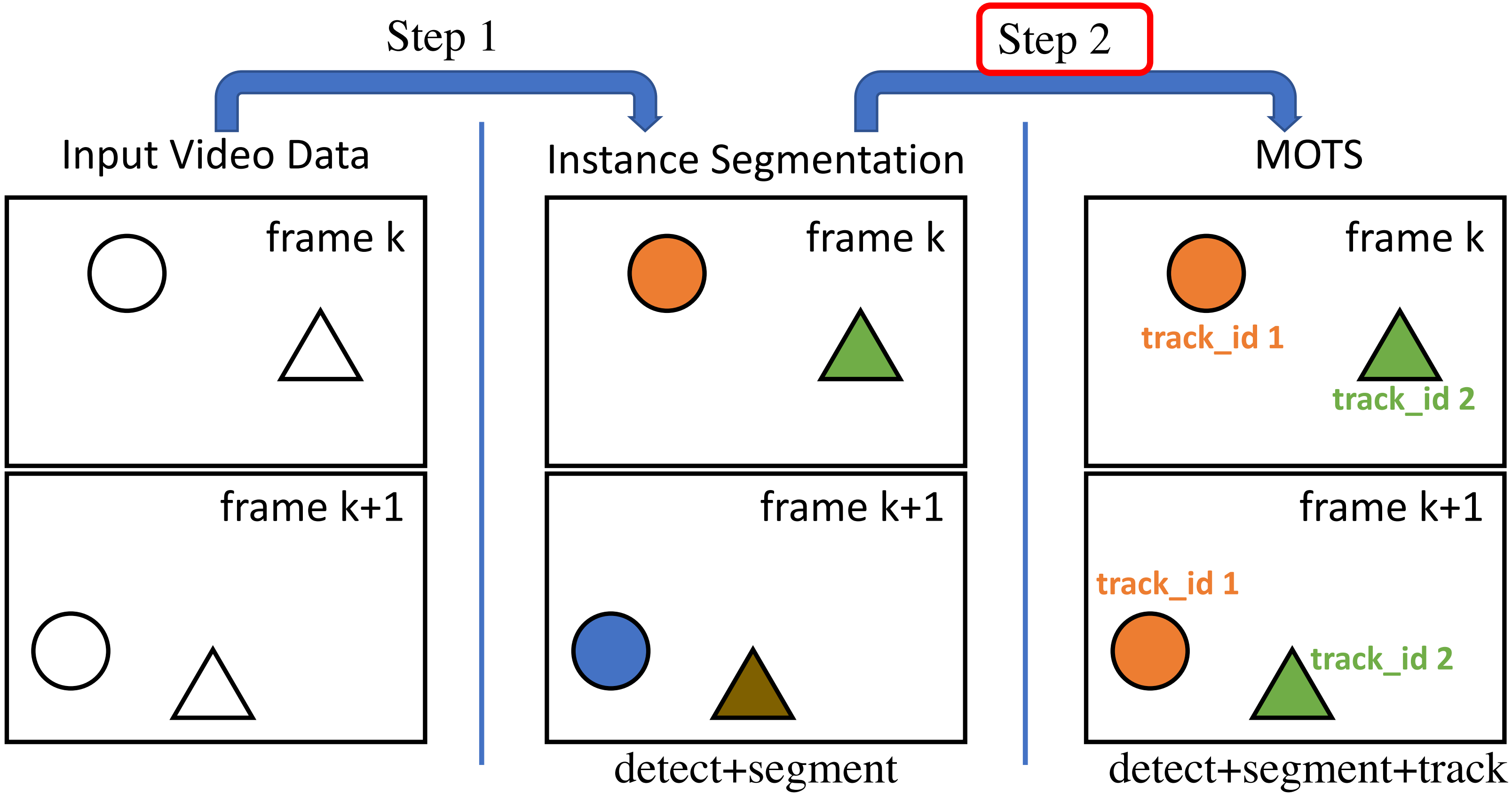


Our solution for MOTS

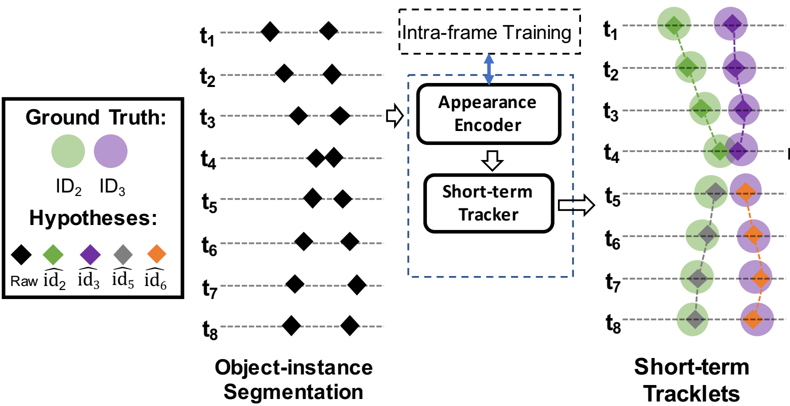
We proposed an offline method, as ReMOTS (Refining Multi-Object Tracking and Segmentation).

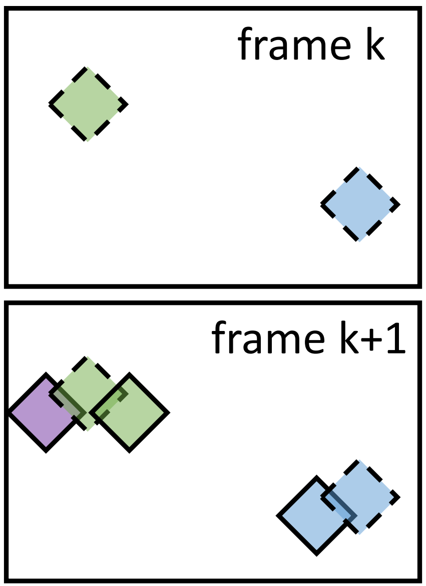
Our main contributions:

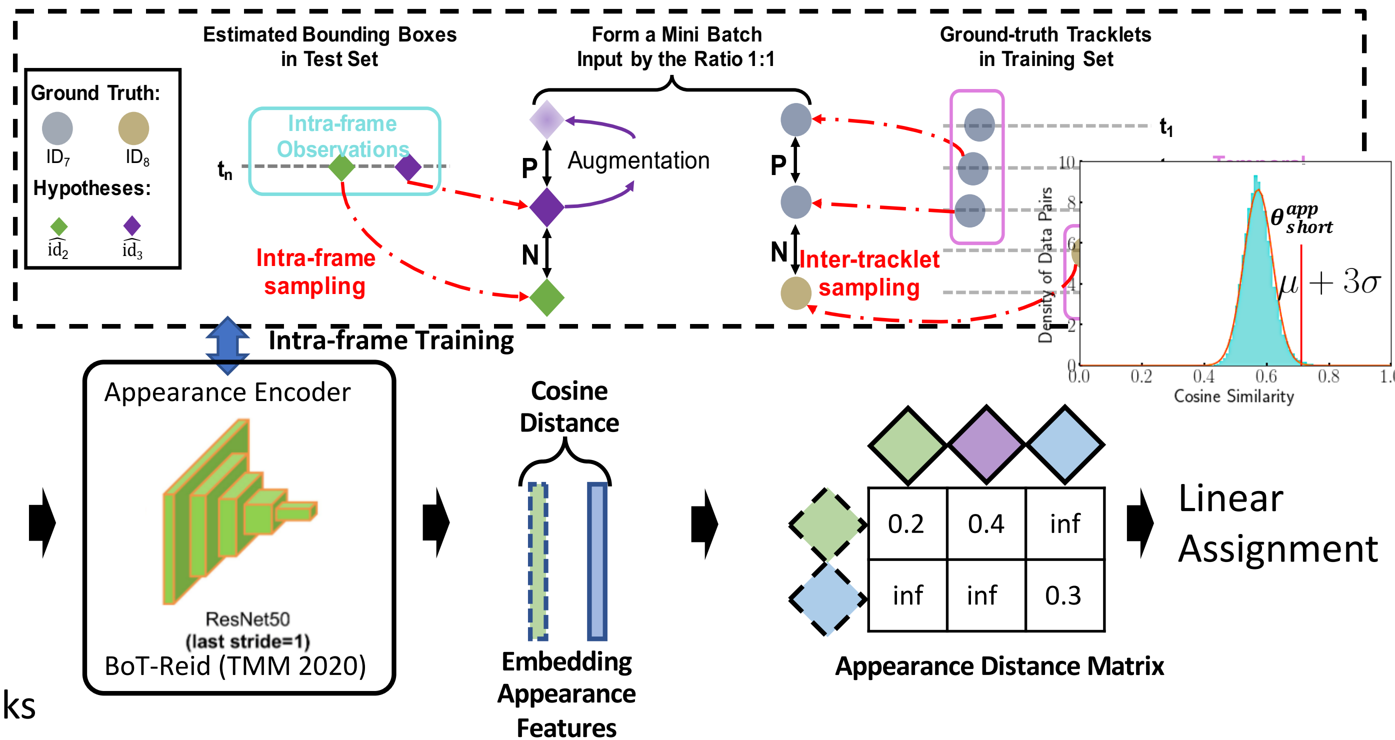
1. Refine appearance features
2. Automatically decide threshold



Intra-frame Training and Short-term Tracking

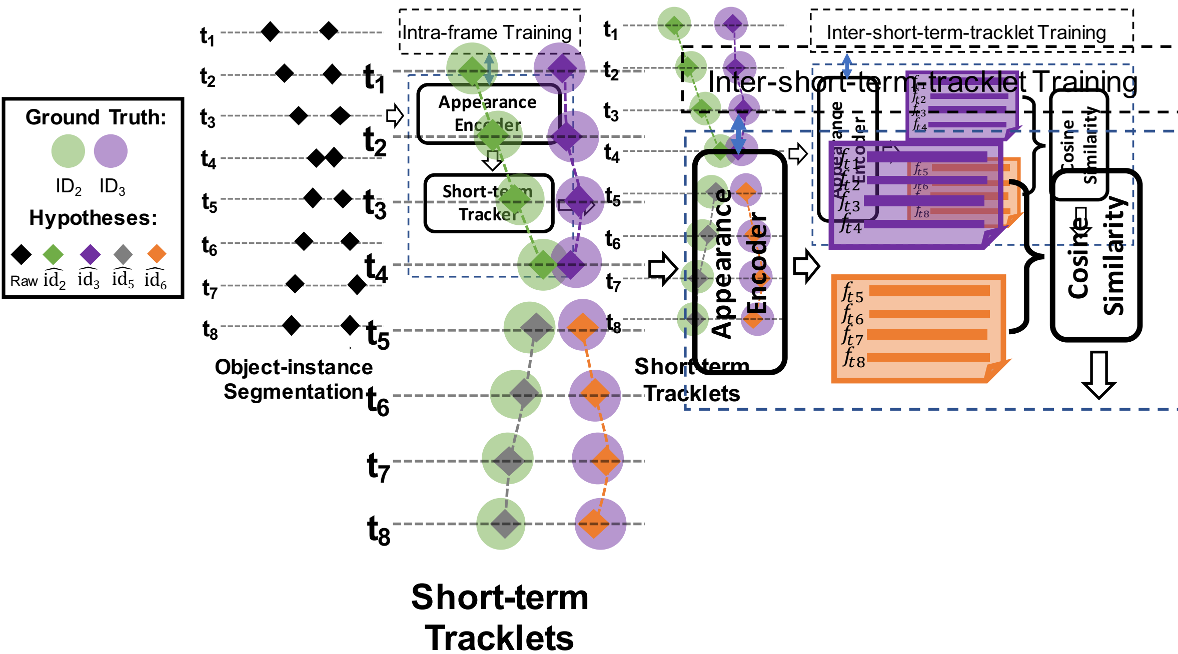


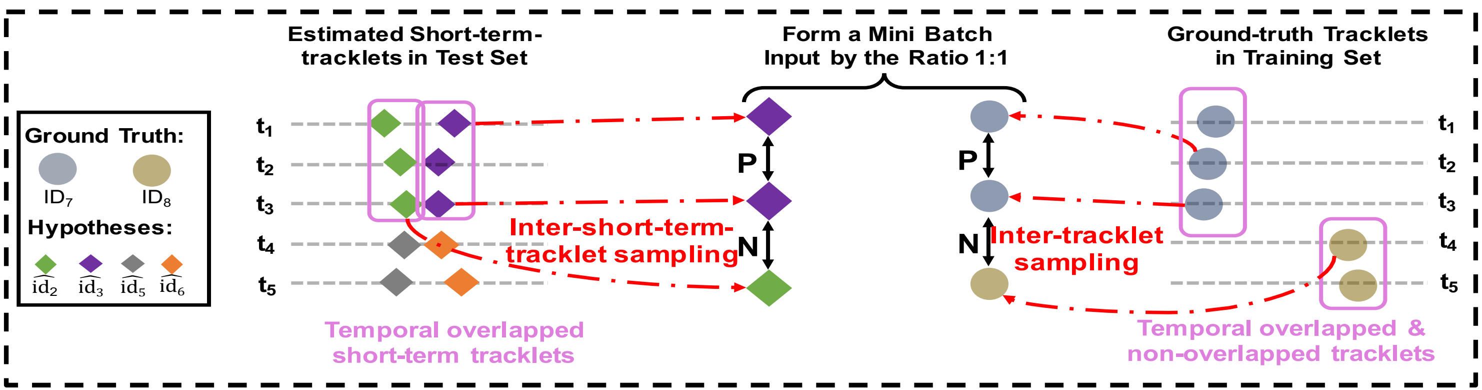




For masks of frame k, consider all of IoU > 0 masks of frame k+1 for matching

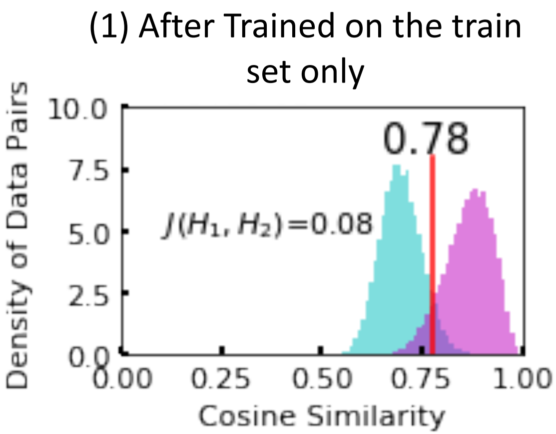
Inter-short-term-tracklet Training

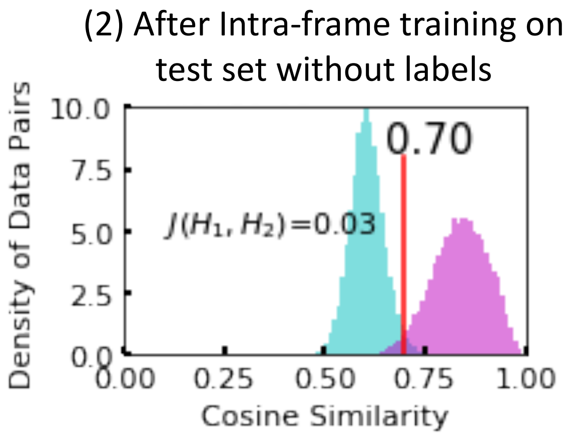


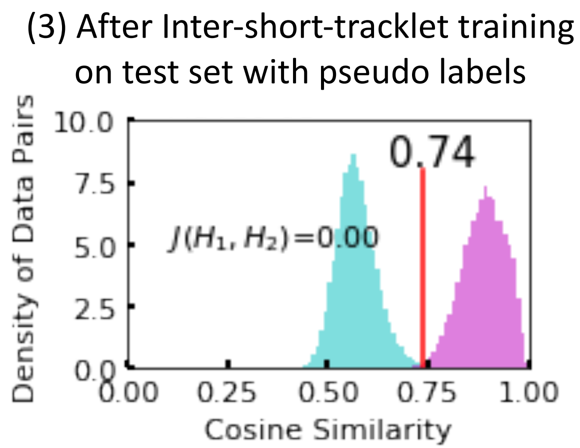


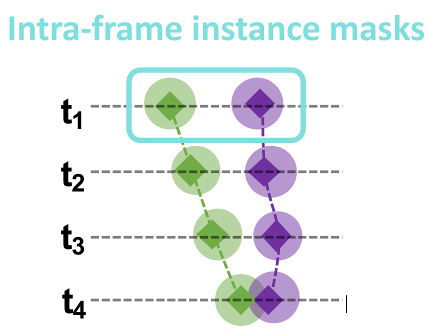
What Happened in Each Step of Appearance Training

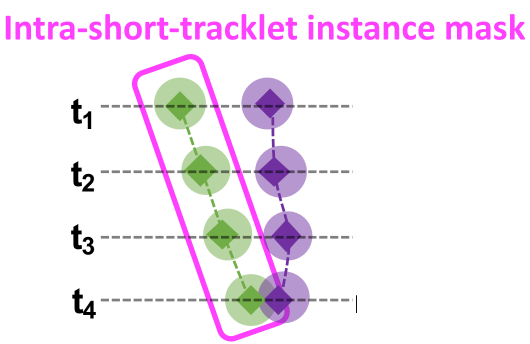
J (H_1, H_2) represents Jaccard Index of two normalized histograms H_1 and H_2.



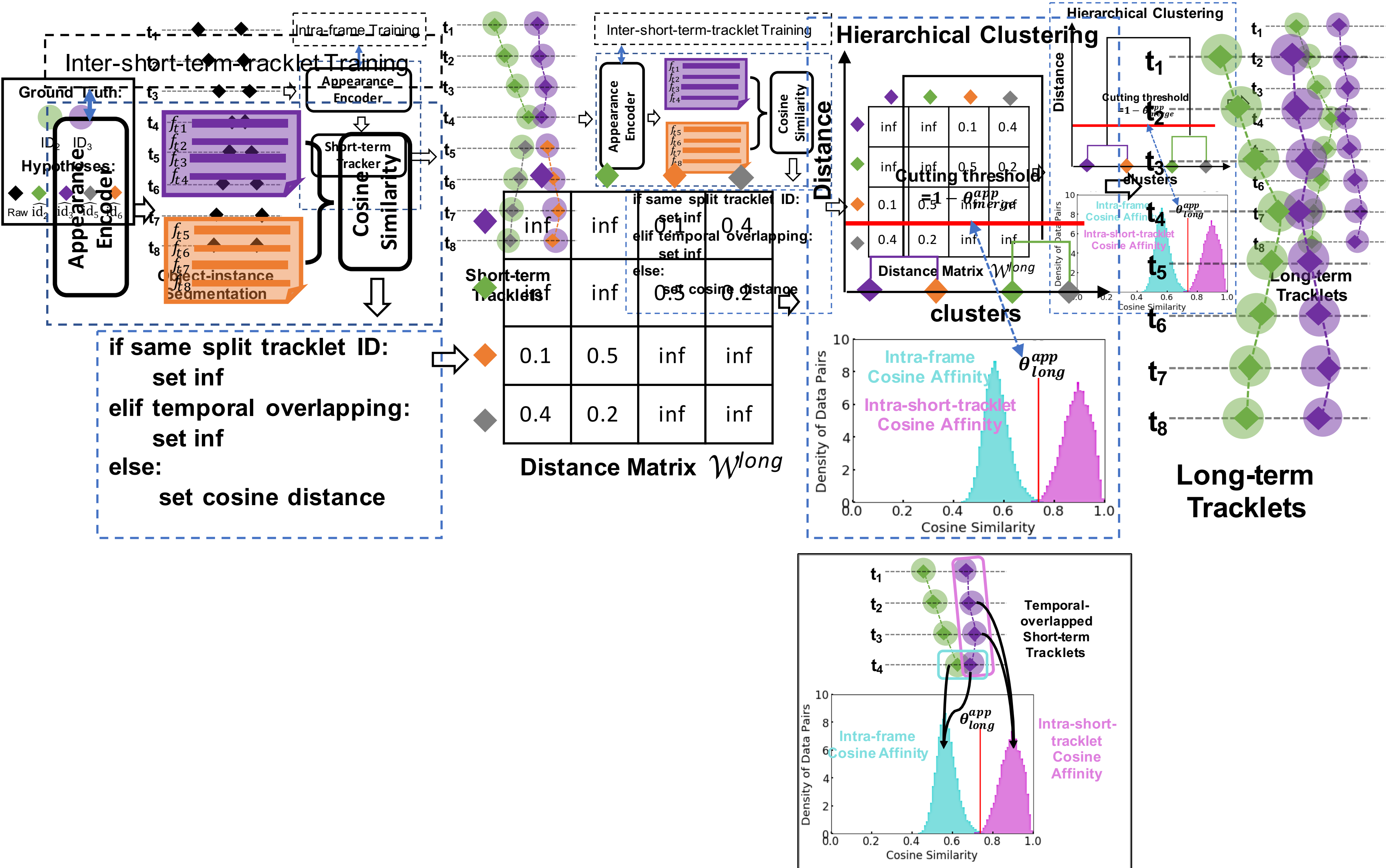








Merging Short-term Tracklets



Comparison with others on MOTSChallenge 1

Benchmark Statistics

Tracker	sMOTSA	IDF1	MOTSA	MOTSP	MODSA	MT	ML	TP	FP	FN	Recall	Precision	ID Sw.	Frag	Hz
ReMOTS 1.	**69.9** ±3.6	75.0 ±5.6	**83.9**	84.0	**85.1**	**248** (75.6)	12 (3.7)	28,270	**819**	3,999	87.6	**97.2**	388 (442.9)	621 (708.8)	**0.3**
														Anonymous submission	
PTPM 2.	68.8 ±3.5	68.5 ±6.2	82.6	84.1	83.7	244 (74.4)	19 (5.8)	28,108	1,084	4,161	87.1	96.3	368 (422.5)	560 (642.9)	10.1
														Anonymous submission	
GMPHD_SAF 3.	68.4 ±3.0	64.9 ±5.5	82.6	83.9	84.4	**248** (75.6)	**10** (3.0)	**28,382**	1,161	**3,887**	**88.0**	96.1	569 (646.9)	770 (875.5)	3.8
														Anonymous submission	
PT 4.	66.8 ±4.9	67.3 ±6.8	79.9	84.5	81.1	234 (71.3)	20 (6.1)	27,215	1,059	5,054	84.3	96.3	370 (438.7)	629 (745.8)	0.4
														Anonymous submission	
DD_Vision 5.	66.6 ±6.2	71.8 ±7.3	79.7	84.4	80.7	243 (74.1)	15 (4.6)	27,114	1,067	5,155	84.0	96.2	341 (405.8)	559 (665.3)	1.6
														Anonymous submission	
Lif_TS 6.	66.3 ±3.4	75.0 ±5.0	79.6	84.2	80.1	224 (68.3)	32 (9.8)	27,112	1,254	5,157	84.0	95.6	**182** (216.6)	525 (624.9)	2.3
														Anonymous submission	
PA 7.	66.2 ±7.1	**76.4** ±5.3	78.9	84.6	79.5	235 (71.6)	21 (6.4)	26,516	849	5,753	82.2	96.9	216 (262.9)	**449** (546.4)	2.5
														Anonymous submission	

May benefit from refinement

May benefit from mask fusion

Since our strategy can be easily adapted to others, will other methods get better performance by applying our appearance encoder and merging?

Limitations of ReMOTS

1. An offline approach.
   - It worth to explore how to bring it to online approach.

2. It is challenging for ReMOTs to handle objects with similar appearance.
   e.g., good for persons (wear different clothes) but not very useful for vehicles (similar textures)

3. Trajectory is not considered in our short-term tracker. Failed to associate fast moving objects.

Slowly moving person with diverse clothes

Fast moving car with similar appearance

Conclusion

- Unlabeled target videos can be used for learning better appearance features, but should take care of the potential of introducing noises.

- The suitable hyper parameters for data association may varies from case to case, and the statistical information of tracklets can be used to adjust them.

- It is preferred to accommodate some insights of ReMOTS to online MOTS.

Thanks for your listening