SynCL: A Synergistic Training Strategy with Instance-Aware Contrastive Learning for End-to-End Multi-Camera 3D Tracking

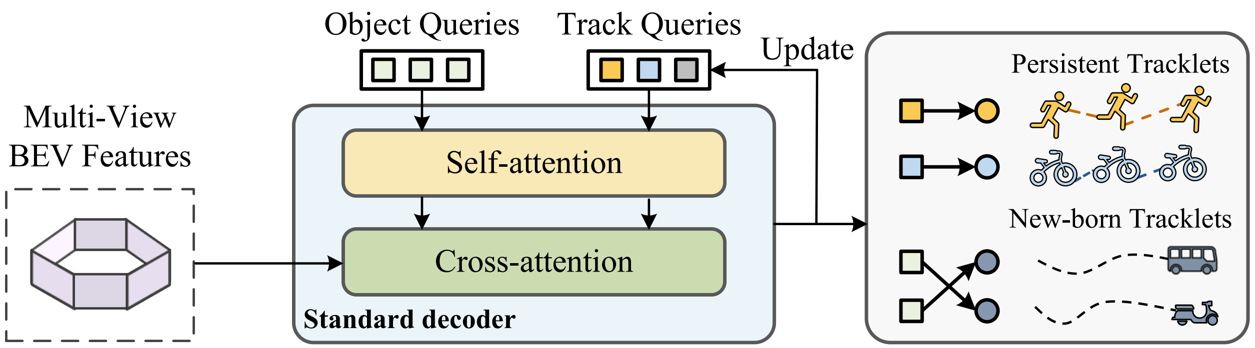
Shubo Lin · Yutong Kou · Zirui wu · Shaoru wang· Bing Li · Weiming Hu · Jin Gao

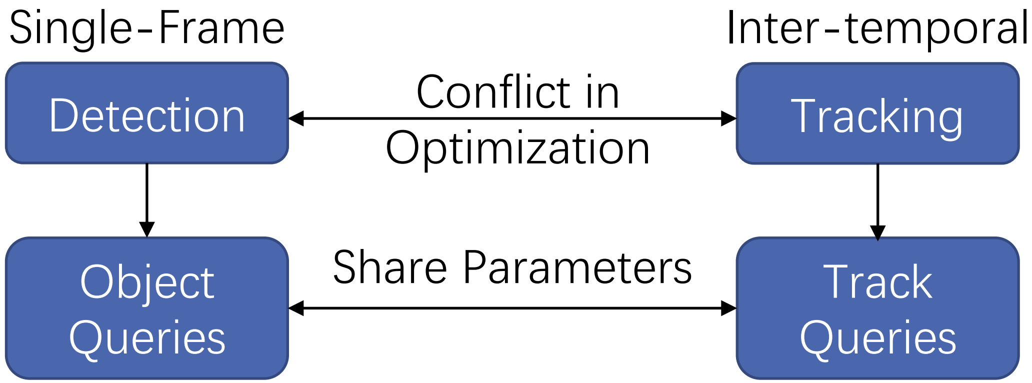
Institute of Automation, Chinese Academy of Sciences

Beijing Key Laboratory of Super Intelligent Security of Multi-Modal Information   People AI

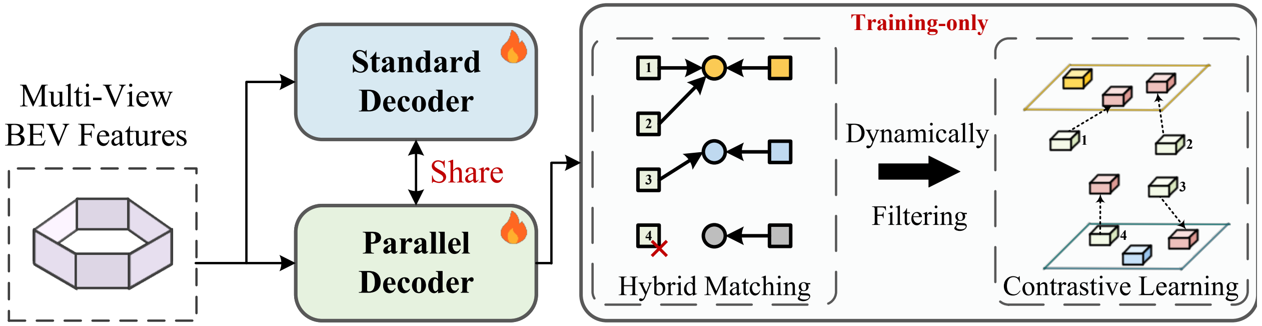
Background and Overview

- Multi-camera Tracking Pipeline



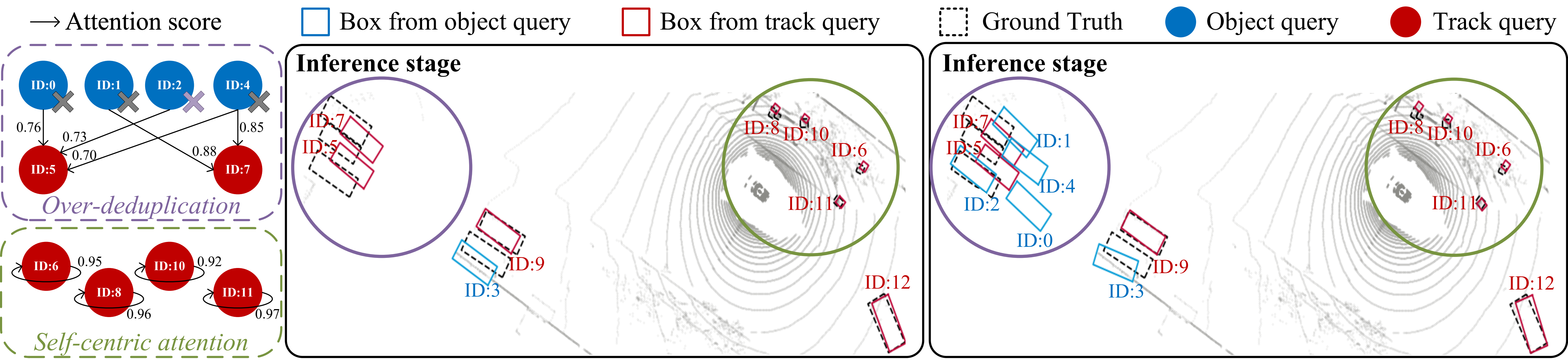


- Our plug-and-play training strategy



Can we address the optimization difficulties by a joint training paradigm without affecting inference speed?

Motivations



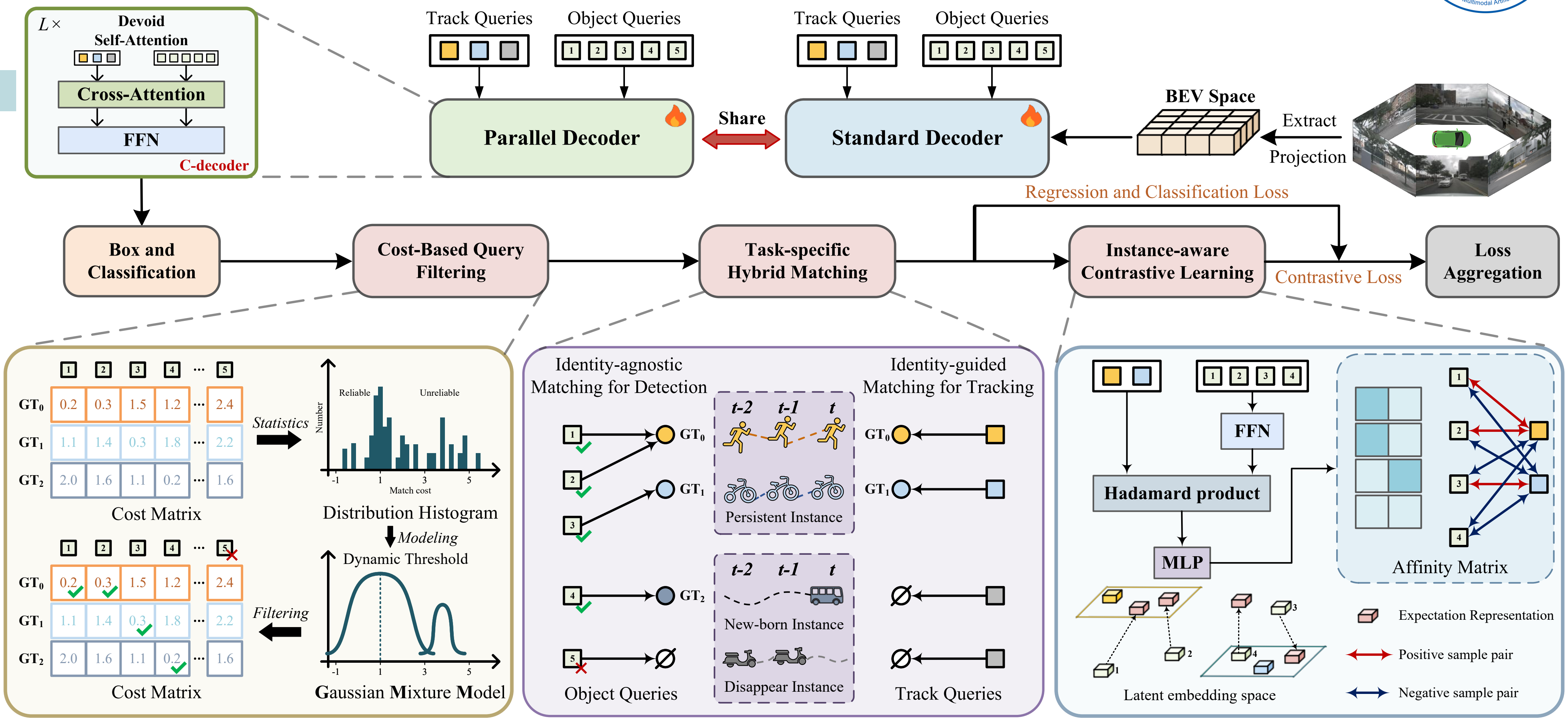
(a) Attention map **(b) Standard decoder w/ Self and Cross Attention** **(c) Remove Self-attention mechanism in Standard decoder**

- Inconspicuous characteristics of self-attention mechanism

Over-deduplication for object queries

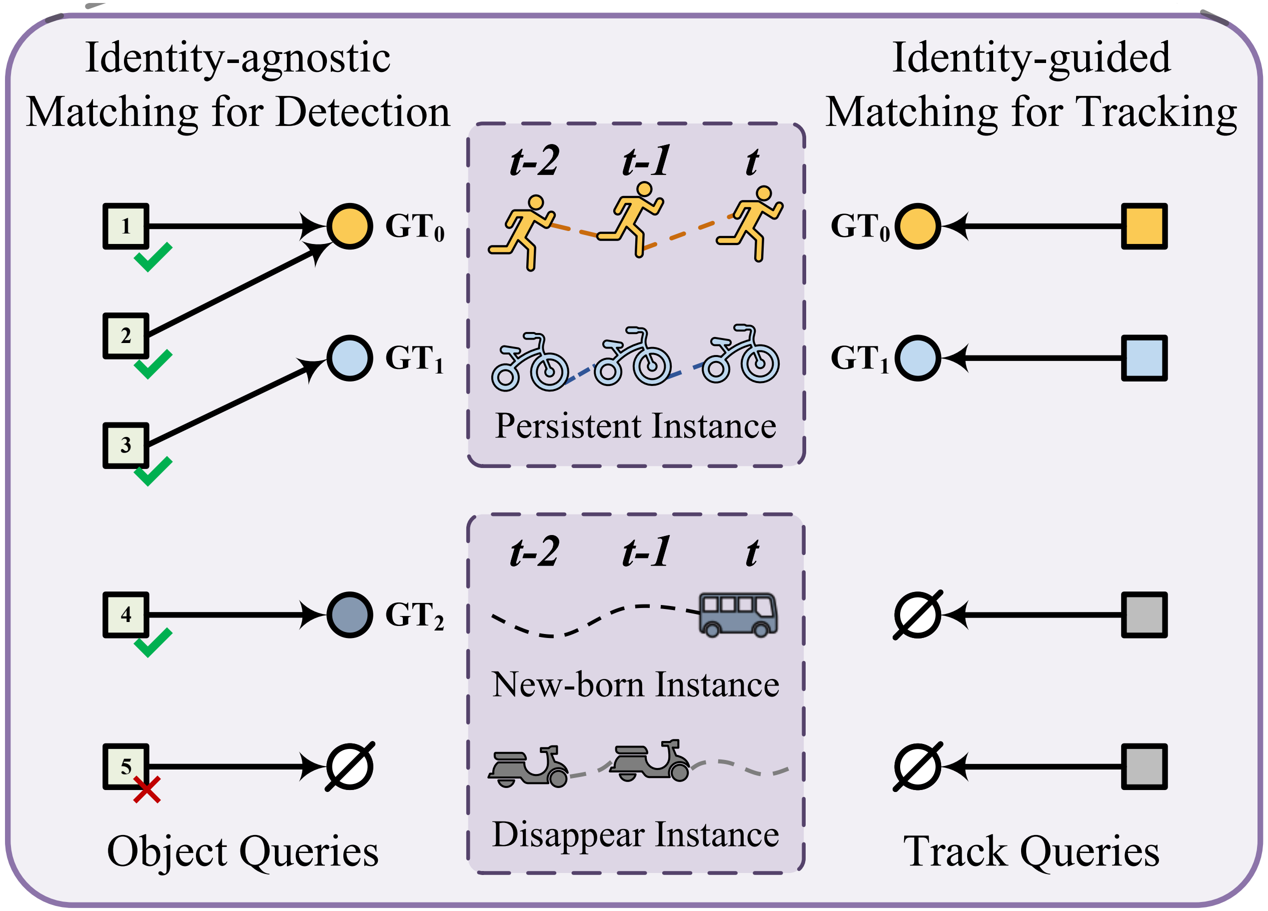
Self-centric attention for track queries

Synergistic Training Strategy



Step 1: Constructing weight-shared parallel decoder w/o self-attention

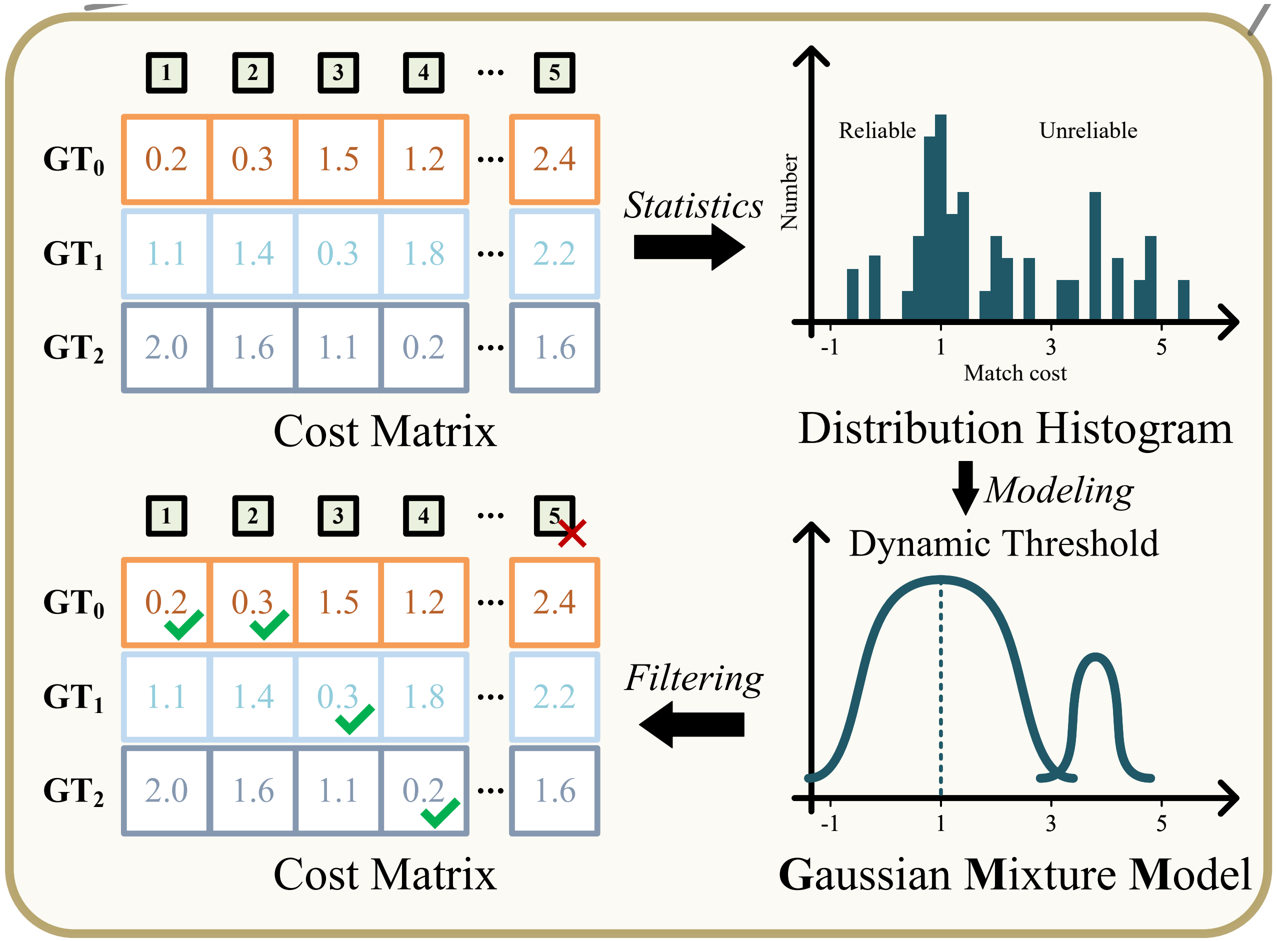
Synergistic Training Strategy



Step 2: Task-specific Hybrid matching in parallel decoder

- One-to-many matching for object queries
  - Recovering high quality candidate boxes overlooked by self-attention
  - Speed up the convergence
- One-to-one matching for track queries
  - Identity-guided ground truth assignment to ensure ID-consistent tracking
  - Background assignment to manage the life cycle of disappear instance

Synergistic Training Strategy

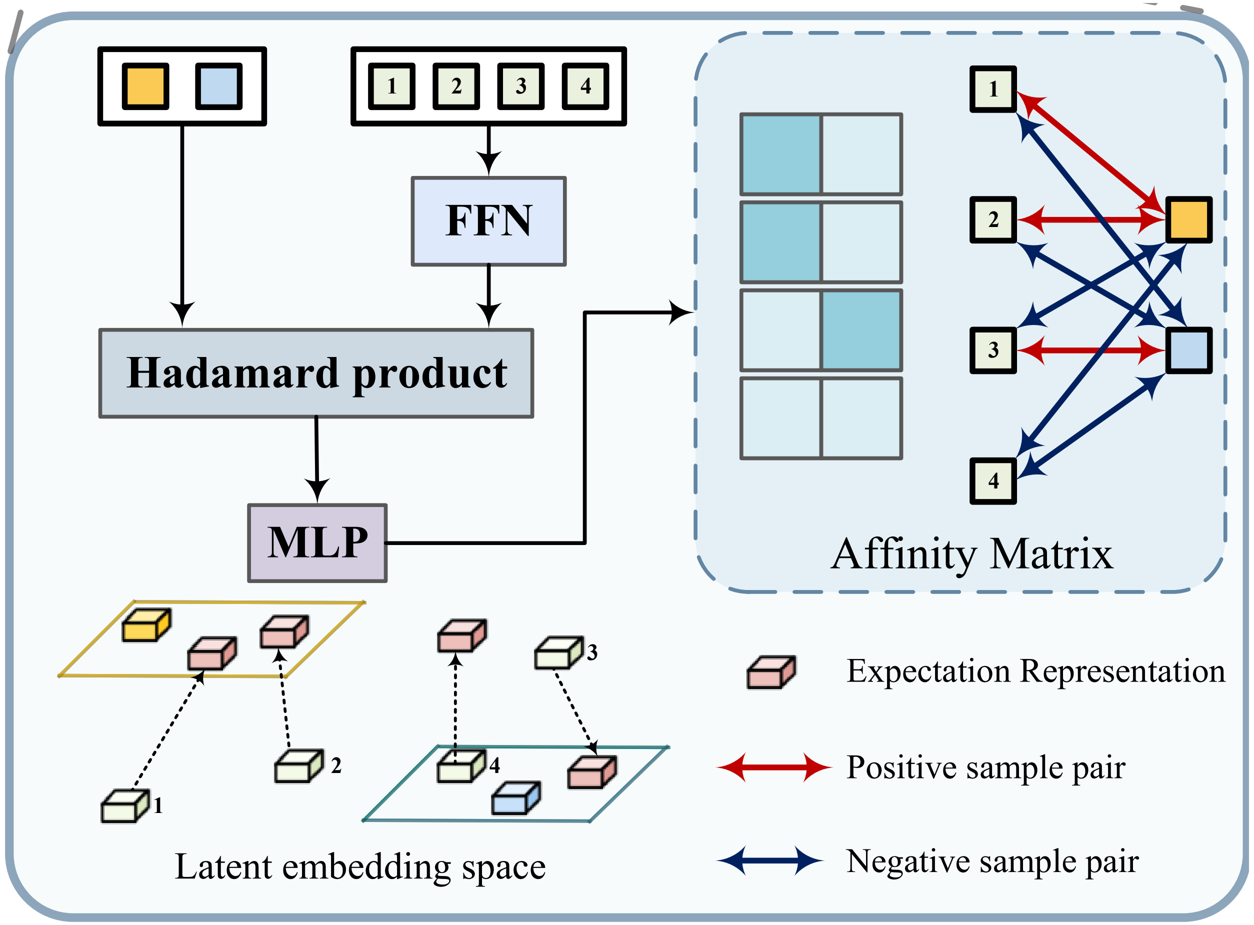


Method	*Tracking*			*Detection*	
	AMOTA	AMOTP↓	Recall	NDS	mAP
IoU-based					
ATSS [41]	39.3%	1.367	49.9%	47.0%	37.3%
SimOTA [42]	42.8%	1.323	52.5%	49.2%	38.8%
Cost-based					
DETA [43]	43.1%	1.320	54.2%	49.4%	38.6%
GMM (ours)	**44.7%**	**1.262**	**56.5%**	**49.7%**	**39.6%**

Dynamic and cost-based method is best for camera-only 3D perception

Step 3: Dynamic Query Filtering for object queries

Synergistic Training Strategy



- Knowledge transfer by representation alignment for object and track queries
- To alleviate the issue of insufficient sample pairs, we utilize kernel-based contrastive learning:

$$\mathcal{L}_{CL}(\mathcal{K};\tau) = -\log \frac{\exp(\mathcal{K}[e_i^{obj}, e_j^{trk}]/\tau)}{\sum_{k=1,k\neq j}^{N} \exp(\mathcal{K}[e_i^{obj}, e_k^{trk}]/\tau)}$$

$$\mathcal{K}[e^{obj}, e^{trk}] = \mathrm{MLP}\left(\left\langle (\mathrm{FFN}(e^{obj}), e^{trk}) \right\rangle_{\mathcal{H}}\right)$$

Step 4: Instance-aware Contrastive Learning for both queries

Results

nuScenes validation set

Method	Backbone	Detector	Resolution	*Tracking*							*Detection*	
				AMOTA	**AMOTP↓**	Recall	MOTA	IDS↓	FP↓	FN↓	**NDS**	**mAP**
MUTR3D* [4]	R101	DETR3D	900 × 1600	32.1%	1.448	45.2%	28.3%	**474**	15269	43828	-	-
SynCL (ours)	R101	DETR3D	900 × 1600	**35.8%**	**1.391**	**49.2%**	**32.9%**	588	**14311**	**40740**	-	-
PF-Track [5]	V2-99	PETR	320 × 800	40.8%	1.343	50.7%	37.6%	**166**	**15288**	40398	47.7%	37.8%
SynCL (ours)	V2-99	PETR	320 × 800	**44.7%**	**1.262**	**56.5%**	**40.8%**	203	15344	**36801**	**49.7%**	**39.6%**
Baseline#1 [5]	V2-99	PETRv2	320 × 800	43.2%	1.272	55.0%	40.6%	**173**	14106	37065	50.4%	41.0%
SynCL (ours)	V2-99	PETRv2	320 × 800	**45.7%**	**1.260**	**56.8%**	**43.0%**	170	**13411**	**36756**	**51.1%**	**42.0%**
Baseline#2 [4]	V2-99	Stream	320 × 800	49.6%	1.164	57.3%	42.9%	**411**	13962	33526	57.6%	48.5%
SynCL (ours)	V2-99	Stream	320 × 800	**51.8%**	**1.149**	**58.8%**	**45.2%**	540	**13639**	**33368**	**58.7%**	**49.2%**

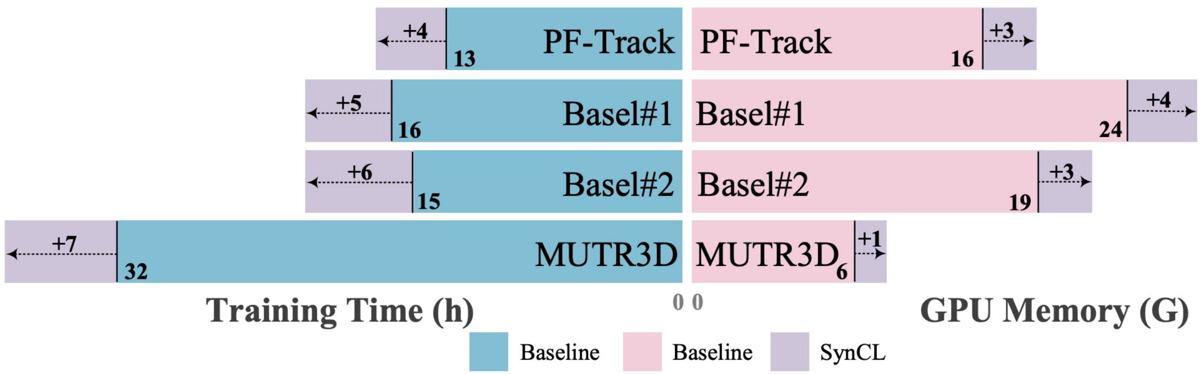
Method	Backbone	Detector	Resolution	**AMOTA**	**AMOTP↓**	Recall	MOTA	IDS↓	FP↓	FN↓	FPS
CC-3DT [39]	R101	BEVFormer	900 × 1600	42.9%	1.257	53.4%	38.5%	2219	-	-	-
DQTrack [11]	V2-99	PETRv2	320 × 800	44.6%	1.251	-	-	1193	-	-	8.6
MUTR3D* [4]	V2-99	PETR	640 × 1600	44.3%	1.299	55.2%	41.6%	**175**	**11943**	36861	6.1
PF-Track [5]	V2-99	PETR	640 × 1600	47.9%	1.227	59.0%	43.5%	181	16149	32778	5.2
ADATrack++ [7]	V2-99	PETR	640 × 1600	50.4%	1.197	60.8%	44.5%	613	14839	30616	3.2
OneTrack [8]	V2-99	Stream	640 × 1600	54.8%	1.088	61.8%	47.9%	389	-	-	-
SynCL (ours)	V2-99	PETR	640 × 1600	50.7%	1.183	61.3%	46.2%	248	14506	30577	5.2
SynCL (ours)	V2-99	Stream	640 × 1600	**58.9%**	**1.016**	**64.0%**	**51.5%**	652	13946	**27330**	5.7

nuScenes test set

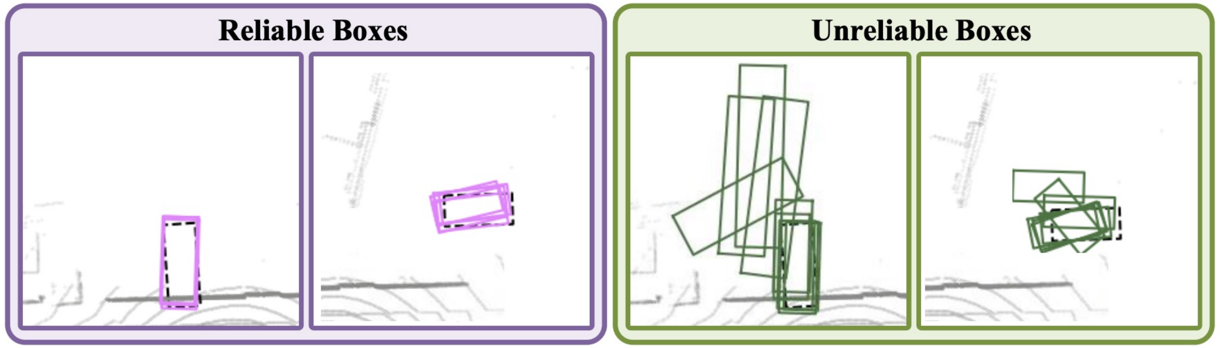
Method	E2E	**AMOTA**	**AMOTP↓**	Recall	MOTA
CC-3DT [39]	✗	41.0%	1.274	53.4%	38.5%
PF-Track [5]	✓	43.4%	1.252	53.8%	37.8%
STAR-Track [24]	✓	43.9%	1.256	56.2%	40.6%
ADATrack++ [7]	✓	50.0%	1.144	59.5%	45.6%
DQTrack [11]	✓	52.3%	1.096	62.2%	44.4%
OneTrack [8]	✓	55.4%	1.021	60.8%	46.1%
DORT [18]	✗	57.6%	**0.951**	63.4%	48.4%
SynCL (ours)	✓	**58.8%**	0.976	**67.1%**	**50.4%**

- Consistent improvements over four baselines with different detectors and tracking frameworks
- Achieving new state-of-the-art performance for the multi-camera 3D MOT task

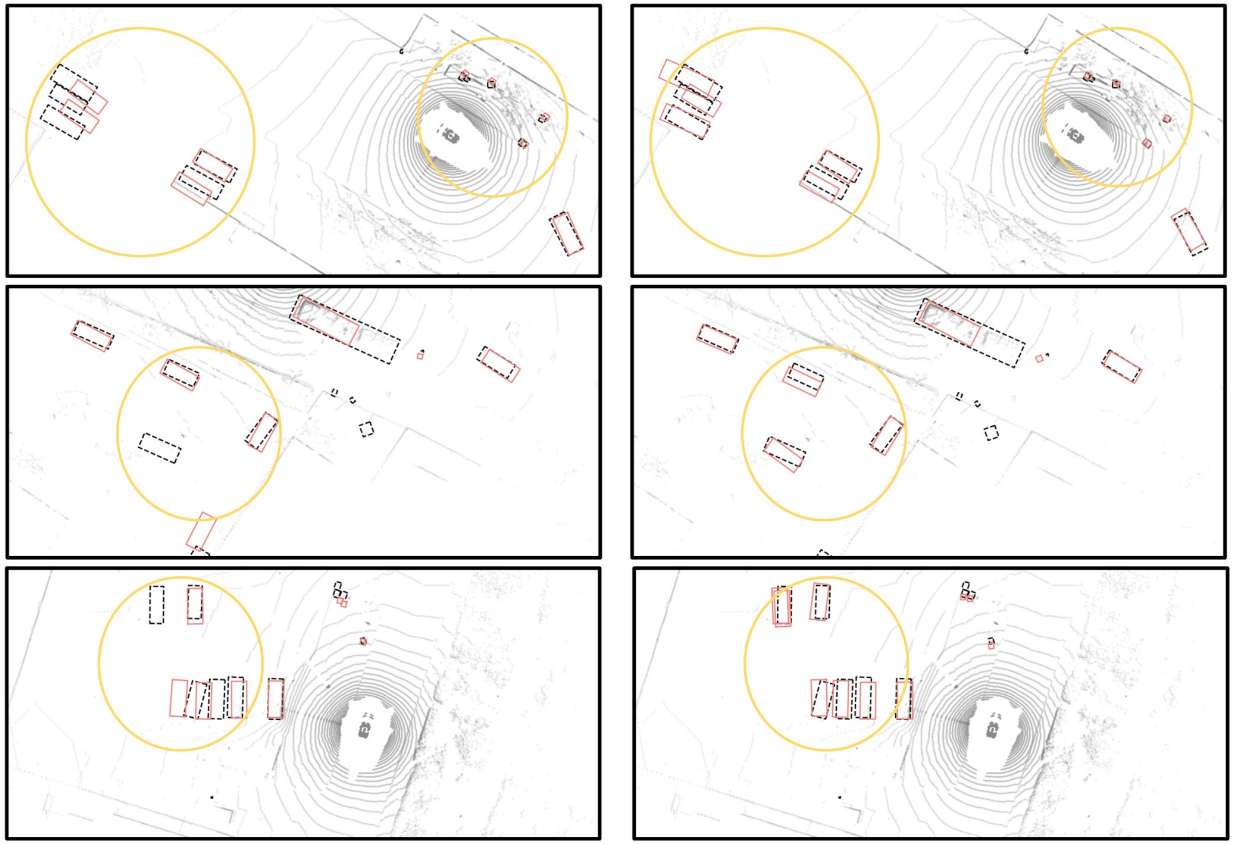
Visualization



- Computation and memory complexity



- Illustration of Dynamic Query Filtering



Original predictions

SynCL (Ours)

Conclusion

Paper

Code

An synergistic training strategy for End-to-End Multi-Camera 3D Tracking

- Reveal the imperceptible effect of the self-attention mechanism across different queries.
- Design a training strategy, implementing dynamic filtering-based hybrid matching and instance-aware contrastive learning.
- Brought remarkable improvements over various tracking-by-attention baselines and achieved new state-of-the-art performance.