MTMMC: A Large-Scale Real-World Multi-Modal Camera Tracking Benchmark

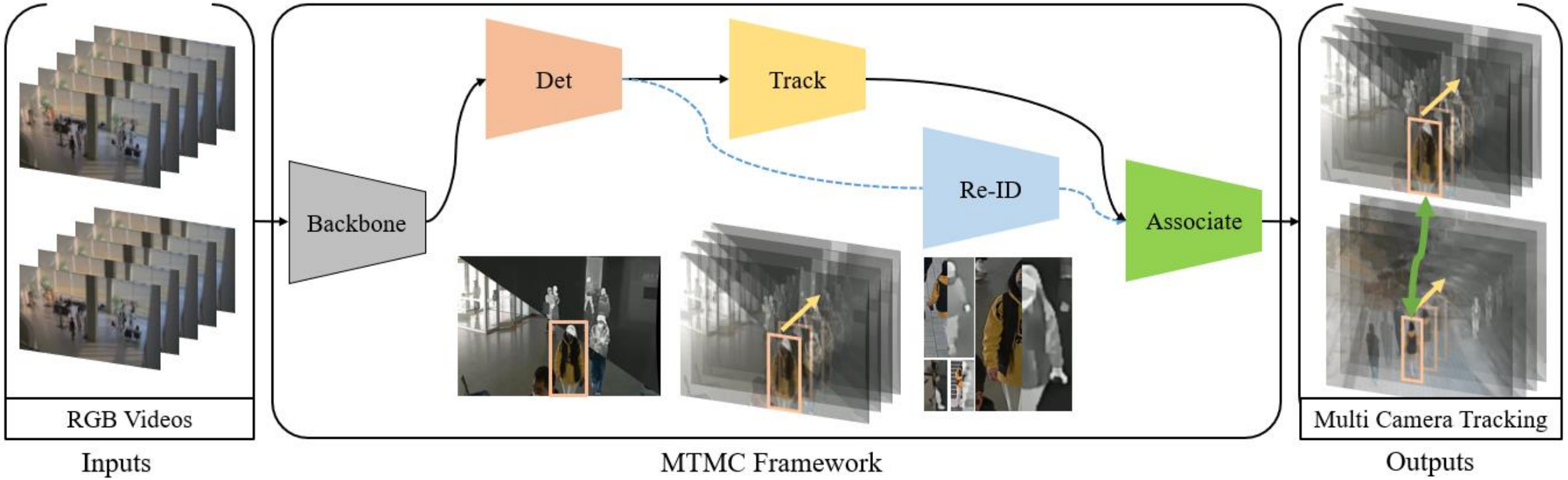
Sanghyun Woo[1*], Kwanyoung Park[2*], Inkyu Shin[3*], Myungchul Kim[3*], In So Kweon[3]
[1]New York University [2]ETRI [3]KAIST (*equal contribution)

Arxiv: https://arxiv.org/pdf/2403.20225

Project: https://sites.google.com/view/mtmmc

Multi-Target Multi-Camera Tracking

:**tracking multiple objects** simultaneously across **different camera views**.



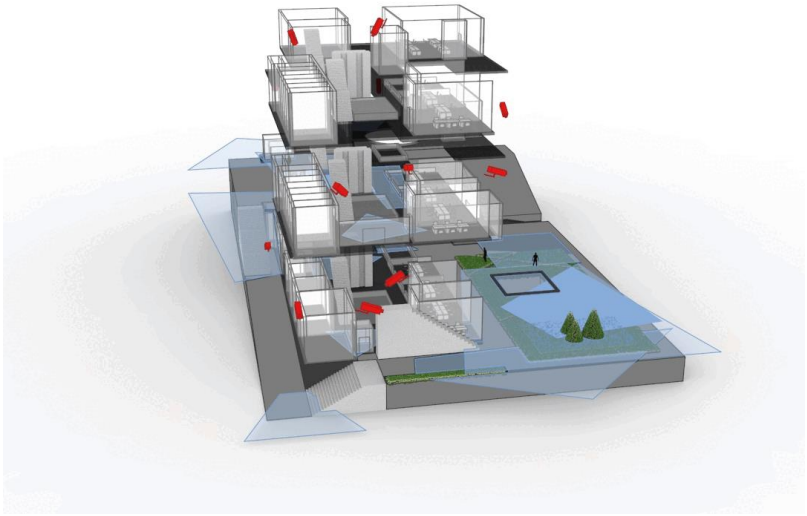
Larger, Longer, and Diverse Tracking Benchmark

Dataset	# Cameras	# ID	# Frames	Camera Coverage	Extra Modality	Resolution
PETS2009	8	30	1,200	outdoor	✗	768 × 576
USC Campus	3	146	135,000	outdoor	✗	852 × 480
Passageway	4	4	120,000	outdoor	✗	320 × 240
NLPR MCT	≤ 5	≤ 235	355,500	in & outdoor	✗	320 × 240
CamNet	8	50	360,000	in & outdoor	✗	640 × 480
WILDTRACK	7	N/A	66,626	outdoor	✗	1920 × 1080
DukeMTMC	8	2,834	2,448,000	outdoor	✗	1920 × 1080
MTA	6	2,840	2,007,360	simulated	✗	1920 × 1080
MMPTRACK	≤ 6	≤ 140	2,979,900	indoor	✗	640 × 320
MTMMC (Ours)	**16**	**3,669**	**3,052,800**	**in & outdoor**	✓ **(Thermal)**	**1920 × 1080**

Multi-modal, Multi-view, Multi-object Videos

Real World Environment

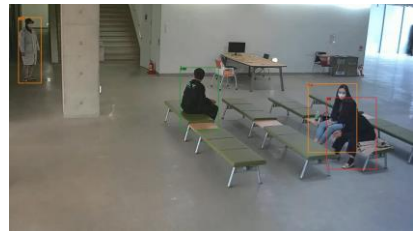
- **Sites:** Campus, Factory

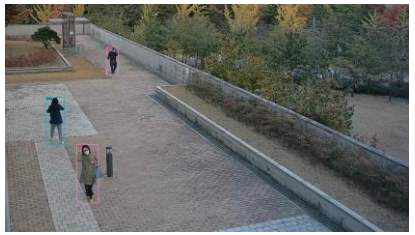



Cam 1


Cam 2


Cam 3


Cam 4

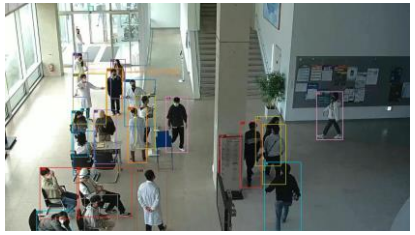

Cam 5


Cam 6


Cam 7

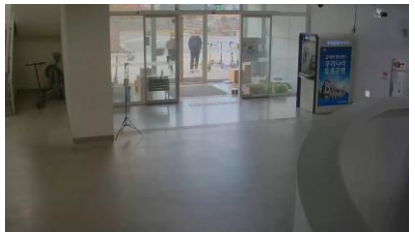

Cam 8

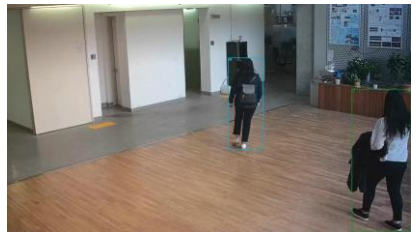

Cam 9


Cam 10


Cam 11


Cam 12


Cam 13

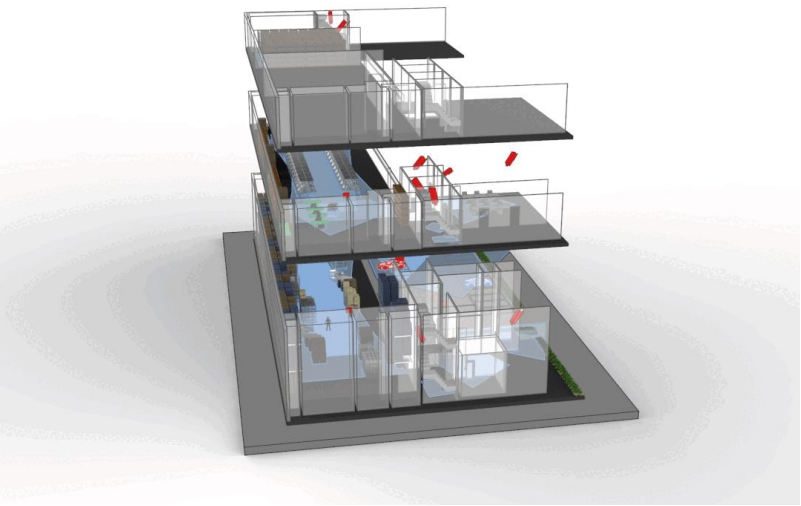

Cam 14


Cam 15

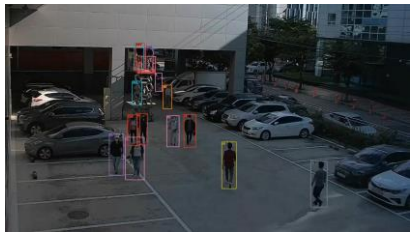

Cam 16

Real World Environment

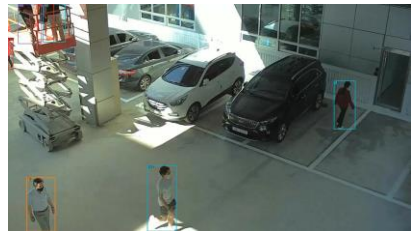
- **Sites:** Campus, Factory



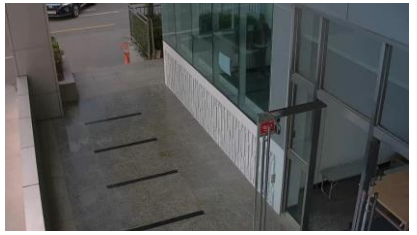

Cam 1


Cam 2


Cam 3


Cam 4


Cam 5

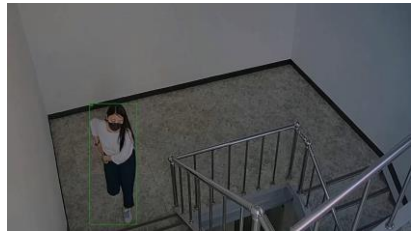

Cam 6


Cam 7


Cam 8


Cam 9


Cam 10


Cam 11


Cam 12


Cam 13


Cam 14

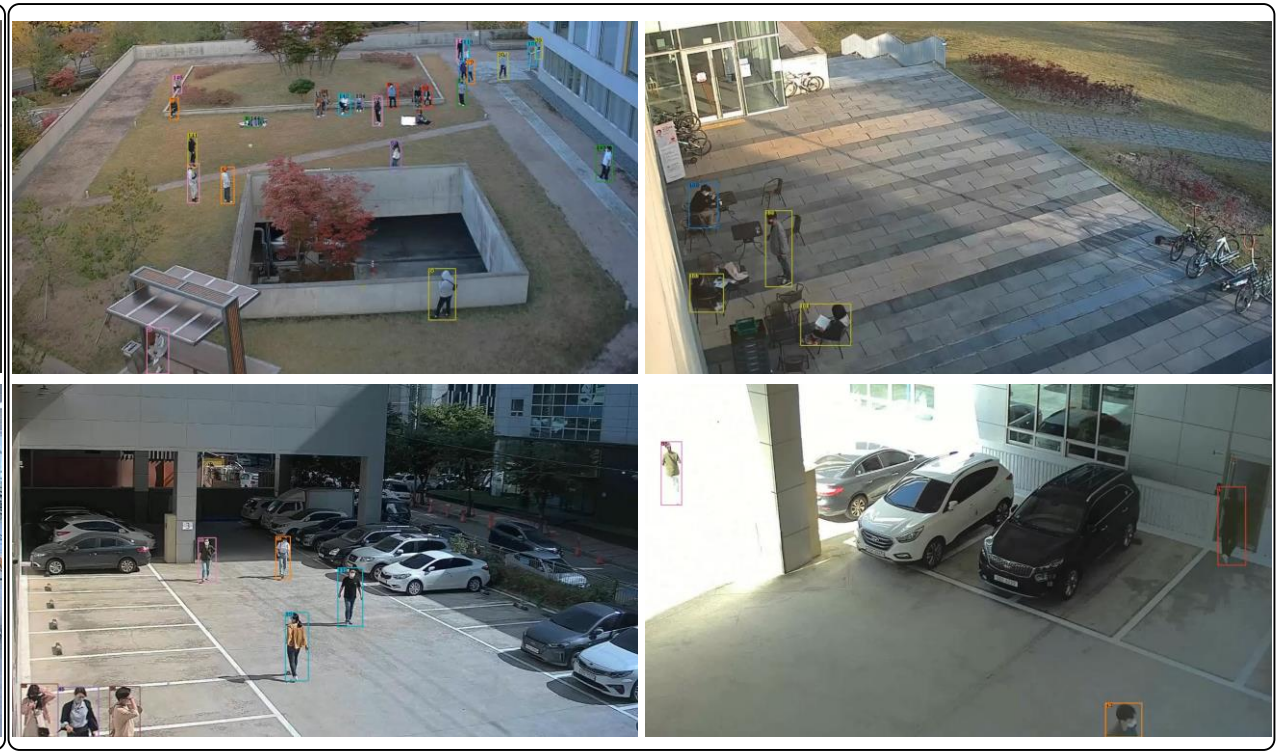

Cam 15


Cam 16

Real World Environment

- **Camera Topology:** Indoor, Outdoor, Multiple Floors, Overlapping Views

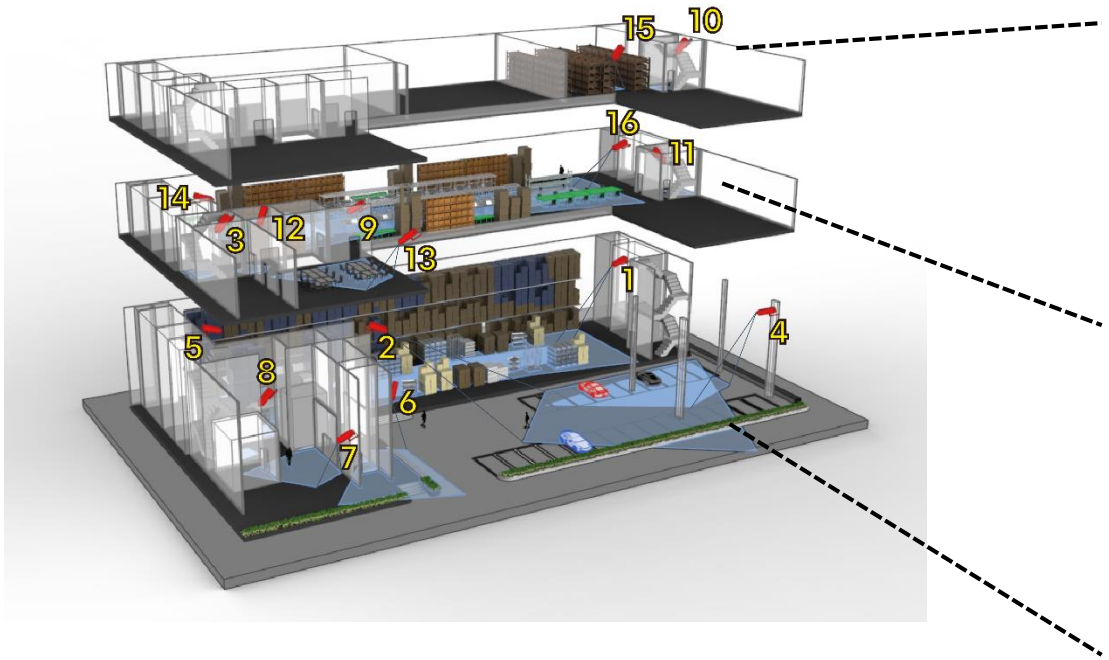
Indoor



Outdoor

Real World Environment

- **Camera Topology:** Indoor, Outdoor, Multiple Floors, Overlapping Views



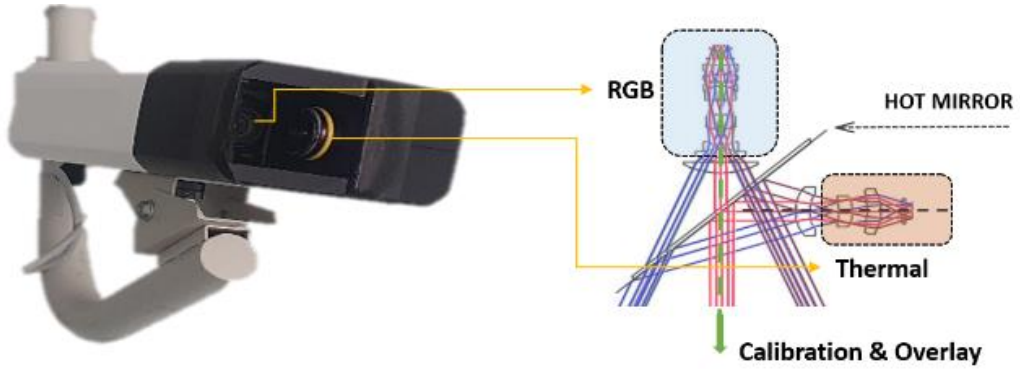
Real World Environment

- **Camera Topology:** Indoor, Outdoor, Multiple Floors, Overlapping Views

1st floor of Factory

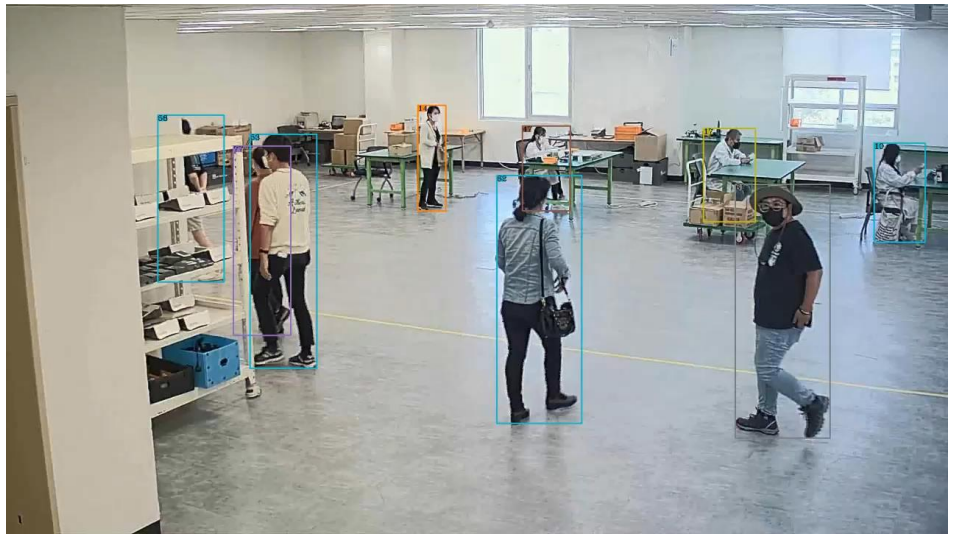
Multi-modal Multi-camera Tracking

- **RGB and Thermal**



RGBT camera with Coaxial Optical System

RGB



Thermal

1. Sub Tasks: Detection, Re-ID

more **challenging** and **generalizable**

Method	Train on	Eval on	mAP
Faster RCNN	COCO-Person	MOT17	29.8
	MTMMC-Person	MOT17	31.3
YOLOX	COCO-Person	MOT17	34.2
	MTMMC-Person	MOT17	38.3

Person Detection

Method	Train on	Eval on	Rank 1	mAP
AGW	Market-1501	Market-1501	95.3	88.2
	MSMT17	MSMT17	78.3	55.6
	MTMMC-reID	MTMMC-reID	76.0	45.6
	MSMT17	Market-1501	64.3	34.2
	MTMMC-reID	Market-1501	66.5	35.4

Person Re-Identification

1. Sub Tasks: Multi-object tracking

more **challenging** and **generalizable**

Method	Train on			Eval on MTMMC					Eval on MOT17				
	MTMMC	MOT17	Misc	IDF1	MOTA	FP	FN	IDs	IDF1	MOTA	FP	FN	IDs
JDE	✓			42.4	74.6	146678	859893	30767	48.0	40.9	2311	29084	329
		✓	cccpe	34.0	52.3	206112	1694301	27347	63.6	60.0	2927	18155	486
	✓	✓	cccpe	43.7	72.6	125770	964863	25725	70.5	65.7	2232	15759	469
QDTrack	✓			53.0	84.5	157529	475242	14542	55.3	43.6	10548	80197	449
		✓		34.3	52.3	286382	1643818	21470	66.8	65.3	9324	45441	1383
	✓	✓		54.2	84.6	439646	439646	14106	70.0	68.6	6927	42903	1005
CenterTrack	✓			50.8	78.6	504642	353525	16972	55.0	45.3	17718	69870	903
		✓		25.2	37.0	629624	1911628	40656	62.1	60.5	6678	55446	1710
		✓	CH_{pre}	27.1	45.7	518692	1662554	40746	63.7	66.2	7128	45939	1611
	✓	✓	CH_{pre}	51.6	80.9	415132	351162	16938	65.7	66.7	6138	46338	1407
ByteTrack	✓			64.8	89.7	112835	300354	7153	69.1	55.9	16896	54106	230
		✓		40.2	56.8	506286	1283368	13585	76.8	75.0	4539	8693	224
		✓	CH	56.9	77.7	267550	640084	7547	79.5	76.6	10128	27250	479
	✓	✓	CH	64.6	89.1	147385	289854	7184	78.7	76.9	8504	28302	517

2. Pre-Training: Real-world vs. Synthetic Data

Powerful pre-trained representations and **synergy** with synthetic data.

Method	Train on		Eval on MOT17				
	MTMMC	MOTSynth	IDF1	MOTA	FP	FN	IDs
QDTrack	✓		55.3	43.6	10548	80197	449
		✓	54.1	43.1	11178	80178	615
	✓	✓	60.8	48.9	14724	67029	870

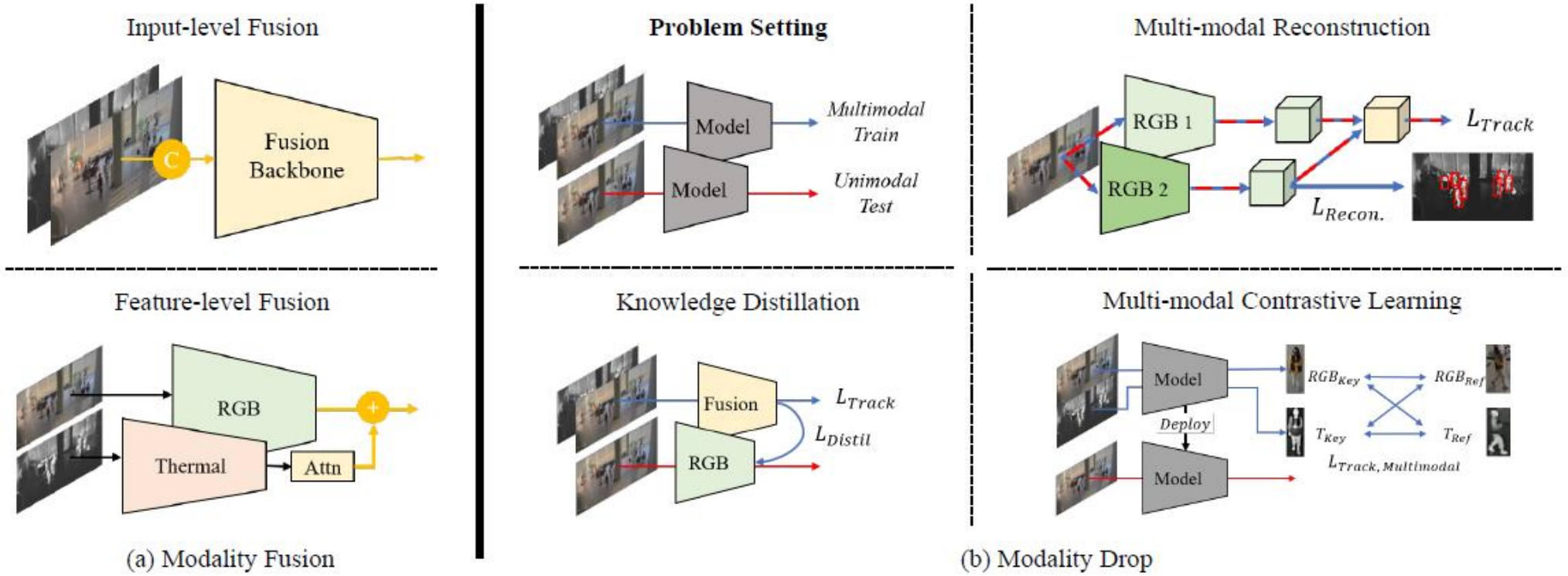
(a) w/o finetune

Method	Train on		Eval on MOT17				
	MTMMC	MOTSynth	IDF1	MOTA	FP	FN	IDs
QDTrack	✓		68.6	66.6	9963	43074	957
		✓	70.8	68.7	9813	39882	921
	✓	✓	72.0	70.2	8367	39135	750

(b) w/ finetune

3. Multi-modal Learning: Setups and Baselines

- **Modality Fusion**: Integrating thermal data either at input or feature level improves performance compared to using single modalities.
- **Modality Drop**: Training with both modalities but evaluating only on RGB shows effective feature transfer and model robustness.



3. Multi-modal Learning: Setups and Baselines

Enhancing Tracking Performance through Modality Fusion.

Method	Fusion	IDF1	MOTA	mAP
RGB	✗	53.0	84.5	92.8
T	✗	44.5	79.2	89.9
RGBT-I	Input	**54.0**	85.6	93.1
RGBT-F	Feature	53.9	**86.0**	**93.5**

(a) Modality Fusion in MTMMC

Method	w/o fine-tune		w/ fine-tune	
	IDF1	MOTA	IDF1	MOTA
RGB-Unimodal (baseline)	55.3	43.6	68.6	66.6
Knowledge Distill.	55.1	43.2	**70.5**	**68.0**
Multi-modal Recon.	57.9	46.2	68.3	67.6
Multi-modal Contrastive.	**59.7**	**48.4**	68.3	67.3

(b) Modality Drop in MTMMC $\rightarrow$ MOT17

4. Multi-modal MTMC

Multi-Target Multi-Camera Tracking Results in MTMMC.

Method	IDF1	MOTA	FP	FN	IDs
TrackTA	32.8	76.9	10604	18715	13364
H. Cluster	41.6	80.9	8012	14663	11072

(a) RGB-based MTMC

Fusion	IDF1	MOTA	FP	FN	IDs
RGBT-I	42.2	81.1	7823	14264	10803
RGBT-F	43.5	81.7	7301	13592	9916

(b) Multi-modal MTMC

Conclusion

- ***The first multi-modal tracking benchmark***
- ***The new multi-modal tracking setups***
- ***The baselines (subtask, multi-modal tracking)***