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

Sanghyun Woo[1*] Kwanyong Park[2*] Inkyu Shin[3*] Myungchul Kim[3*] In So Kweon[3]

[1]New York University [2]ETRI [3]KAIST

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

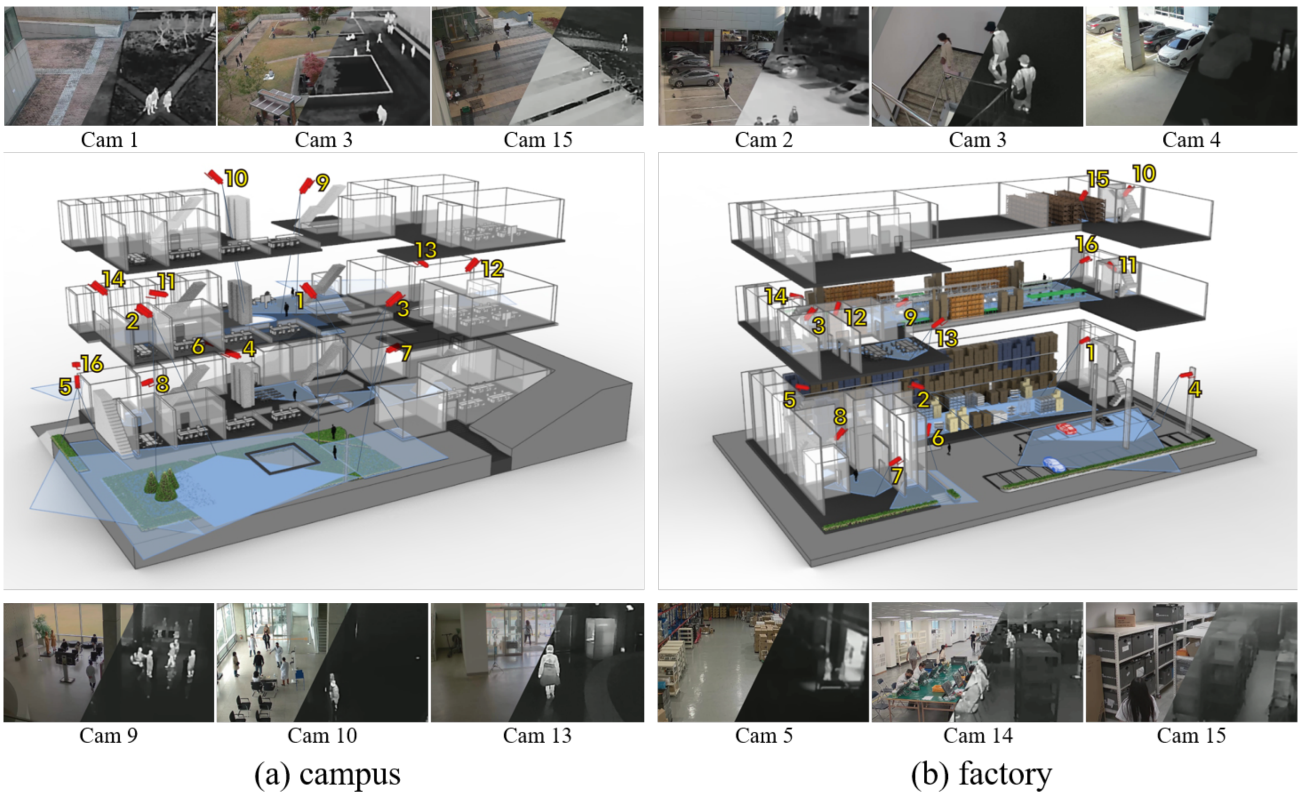

Figure 1. **The 3D layout overview.** (a) campus and (b) factory. We installed 16 multi-modal cameras in both indoor and outdoor settings, across multiple floors, with overlapping coverage. The cameras were fixed in position and angle to densely cover the building, creating a realistic surveillance camera system.

Abstract

Multi-target multi-camera tracking is a crucial task that involves identifying and tracking individuals over time using video streams from multiple cameras. This task has practical applications in various fields, such as visual surveillance, crowd behavior analysis, and anomaly detection. However, due to the difficulty and cost of collecting and labeling data, existing datasets for this task are either synthetically generated or artificially constructed within a controlled camera network setting, which limits their ability to model real-world dynamics and generalize to diverse camera configurations. To address this issue, we present MTMMC, a real-world, large-scale dataset that includes long video sequences captured by 16 multi-modal cameras in two different environments - campus and factory - across various time, weather, and season conditions. This dataset provides a challenging test-bed for studying multi-camera tracking under diverse real-world complexities and includes an additional input modality of spatially aligned and temporally synchronized RGB and thermal cameras, which enhances the accuracy of multi-camera tracking. MTMMC is a super-set of existing datasets, benefiting independent fields such as person detection, re-identification, and multiple object tracking. We provide baselines and new learning setups on this dataset and set the reference scores for future studies. The datasets, models, and test server will be made publicly available.

* Equal contribution

1. Introduction

Multiple object tracking (MOT) is an essential vision task that helps us understand visual content and predict the evolution of the surroundings over time. Recent advancements in MOT, thanks to benchmarks such as MOT17 [39], BDD100K [68], Waymo [51], and TAO [10] have led to the development of more effective and efficient trackers [16, 38, 40, 61, 82]. Despite these advancements, multiple-camera tracking has seen limited exploration, largely due to the lack of appropriate datasets. The high costs associated with the collection and annotation of such data are a major bottleneck.

The datasets currently available predominantly consist of either synthetically generated data from game simulators [27] or small-scale real-world data obtained from controlled camera networks [3, 6, 11, 13, 17, 18], which assume an idealized overlap between the camera views to simplify the annotation process. However, synthetic data often fail to translate effectively to real-world scenarios due to significant domain shifts, and datasets from controlled environments do not reflect the complexities of real-world multi-camera networks. Additionally, the withdrawal of the DukeMTMC [46], previously the most extensive real-world dataset, due to privacy issues has left a considerable void in this research area.

To tackle this, this paper presents a new benchmark called the Multi-Target Multi-Modal Camera (MTMMC) tracking dataset. The dataset was collected from two challenging environments—a campus and a factory—equipped with 16 multi-modal cameras, each placed at different angles (see Fig. 1). The dataset consists of 25 video recordings—13 from the campus and 12 from the factory—with each video containing five and a half minutes of HD video recording captured under various times, weather, and seasons, ensuring a rich diversity of backgrounds. To ensure compliance with data privacy standards, we collected informed consent from all participants, who explicitly agreed to the public release of the collected data for research purposes. The annotation of all trajectories was accomplished using a semi-automatic labeling system, carefully refined by crowdworkers over *a year*, making the dataset the most largest publicly accessible MTMC tracking benchmark to date.

Significantly, our dataset contains both RGB and thermal cameras, allowing the tracker to additionally utilize thermal information for more accurate multi-camera tracking. This is the first time a dataset has provided a valid test-bed for studying the impact of multi-modal learning for multi-camera tracking. Our experiments reveal that incorporating thermal data into standard RGB camera-based trackers results in more robust tracking, motivating future research in this new direction. The construction of the MTMC dataset also facilitates progress in related subtasks, such as person detection, re-identification, and MOT.

2. Related Work

Benchmarks To construct a high-quality MTMC dataset, it is crucial to have temporally synchronized videos from multiple cameras. These videos must also maintain consistent person identities across all camera views. However, this requirement results in high annotation costs. As a result, existing MTMC benchmarks are either short in duration [11, 13, 17, 18], have low video resolution [3, 6, 29, 72], or provide inconsistent person IDs [7], making them unsuitable for training generic deep trackers for real-world use cases. The most popular dataset closest to our proposal is DukeMTMC [46], but it was withdrawn due to consent and privacy issues. Recently, two large-scale MTMC datasets, MTA [27] and MMPTRACK [20], have been introduced, but they have limitations such as being obtained through game simulations or collected in controlled setups where all cameras have overlapping fields of view. The new MTMC dataset aims to provide a larger basis for training and testing MTMC performance than any previous datasets, making it a valuable resource for researchers.

Multi-modal Learning Unlike existing tracking datasets, our dataset features an additional thermal input modality, which opens up new research directions for multi-modal learning in multi-camera tracking. Multi-modal learning is an interesting research problem that not only improves model robustness through modality fusion, which applies to various vision tasks such as detection [31, 70, 71, 71, 71], visual object tracking [25, 26, 28, 35, 65], and segmentation [53, 81], but also enables better representations of each modality by learning the intrinsic correlations between them [1, 37, 48, 64, 76]. In this paper, we present two new experimental setups with baselines.

Multiple Object Tracking The standard way to tackle the MTMC problem involves a two-step approach: 1) generating local tracklets for all the targets within each camera; 2) associating these local tracklets across cameras when they belong to the same target. The first step, known as multiple object tracking, has been extensively studied by the community. The tracking-by-detection paradigm has emerged as the dominant approach, owing to significant improvements in object detection techniques [33, 34, 43, 44]. Recent advances in this paradigm include developing more discriminative association objectives [23, 30, 40, 41, 55, 59, 60], unifying detection and tracking [16, 52, 62, 82], or building an end-to-end framework [38, 50, 69].

Multi Camera Association Cross-camera association presents a more challenge due to pronounced changes in object appearance between cameras, variable background conditions, and an increased number of targets to be matched. To facilitate this process, various constraints have

Dataset	# Cameras	# ID	# Frames	OV/NOV	Camera Coverage	Extra Modality	FPS	Resolution
PETS2009 [17]	8	30	1,200	OV	outdoor	✗	30	768×576
USC Campus [29]	3	146	135,000	NOV	outdoor	✗	30	852×480
Passageway [3]	4	4	120,000	OV	outdoor	✗	25	320×240
NLPR MCT [6]	≤ 5	≤ 235	355,500	NOV	in & outdoor	✗	20	320×240
CamNet [72]	8	50	360,000	NOV	in & outdoor	✗	25	640×480
WILDTRACK [7]	7	N/A	66,626	both	outdoor	✗	60	1920×1080
DukeMTMC [46]	8	2,834	2,448,000	NOV	outdoor	✗	60	1920×1080
MTA [27]	6	2,840	2,007,360	both	simulated	✗	41	1920×1080
MMPTRACK [20]	≤ 6	≤ 140	2,979,900	OV	indoor	✗	15	640×320
MTMMC (Ours)	**16**	**3,669**	**3,052,800**	**both**	**in & outdoor**	✓ **(Thermal)**	23	**1920 ×1080**

Table 1. **Overview of the publicly available MTMC datasets.** For each dataset, we report the number of cameras, person identities, and frames. We also report the presence of overlapping (OV) / non-overlapping (NOV) camera views, camera coverage, availability of extra input modality, annotated frame rate (FPS), and frame resolution. Our new MTMMC dataset is unprecedented in its scale and diversity. It includes 16 cameras, 3,669 person IDs, and 3 million frames, making it a challenging and large-scale dataset. The dataset also provides high-resolution and multi-modal information.

been employed, including time conflicts [75], linear motion patterns [45], camera network topology [24, 49], geometric cues [5, 8, 67], and spatial locality [22].

Cross-camera association can be conducted in both real-time [6, 42] and offline manners [9, 21, 22, 45, 75], with the latter often favored for its enhanced accuracy. Notably, several offline global association techniques have been developed, such as hierarchical clustering [75, 80], correlation clustering [4, 45], matrix factorization [21], and adaptive locality-based association [22].

3. MTMMC

Our camera network is detailed in Fig. 1, comprising a school campus and a factory environment. This choice reflects common real-world surveillance scenarios. With 16 multi-modal cameras, our configuration spans both indoors and outdoors, extends over multiple levels of floors. Each camera provides RGB and thermal data that are spatially aligned and temporally synchronized. A detailed description of all cameras is provided in the appendix.

3.1. Data Characteristics

Table. 1 presents a comparative summary between our dataset and the existing datasets. The MTMMC dataset consists of 25 scenarios, each composed of 16 **high-resolution** RGB + Thermal videos captured at 23 fps, both at **indoor and outdoor**, resulting in a total of 3,052,800 frames. The dataset offers diverse real-world environmental conditions, ranging from day to evening (*time*), sunny to cloudy (*weather*), and summer to fall (*season*). This **diversity** makes our dataset unique and more representative.

In Fig. 2, we present a detailed comparison of our MTMMC dataset with two of the largest datasets in MTMC: MTA [27] and MMPTrack [20]. We focus on three critical aspects that influence the tracking performance:

1. **Track Length.** We analyze the variability in the number of cameras each individual is tracked through, providing insights into the robustness of tracking across cameras. MTMMC features a broader distribution of track lengths, including more extended tracking periods compared to MTA and MMPTrack, as shown in Figure 2(a), demonstrating our dataset's capability to *challenge and train models on maintaining identities over longer sequences*.
2. **Track Scale.** We assess the range of target scales by normalizing bounding boxes against image dimensions. MTMMC includes a diverse range of scales, with Figure 2(b) highlighting instances of small-scale tracks that are not well-represented in other datasets, critical for *training models to detect and track small or distant targets*.
3. **Track Path.** Our analysis of track trajectories in normalized image coordinates reveals MTMMC's coverage of diverse movement patterns. The dataset exhibits a more comprehensive range of trajectories compared to others, as depicted in Figure 2(c), which is pivotal for algorithms *predicting and maintaining track continuity amidst complex environmental dynamics*.

MTMMC advances in all these three aspects over the previous datasets, providing a challenging testbed that more precisely reflects real-world conditions.

We further provide a statistical analysis of key attributes of our dataset in the appendix. First, the number of objects per frame and the number of tracks per video metrics reveal the **complexity** of scene contexts and the robustness required for successful multi-object tracking. Additionally, the **wide demographic range**, represented by the age and gender distributions of the actors ensures the development of more inclusive and unbiased tracking algorithms. Lastly, for the first time, a thermal modality is included, enabling **multi-modal** learning — a feature unprecedented in previous multi-object tracking datasets.

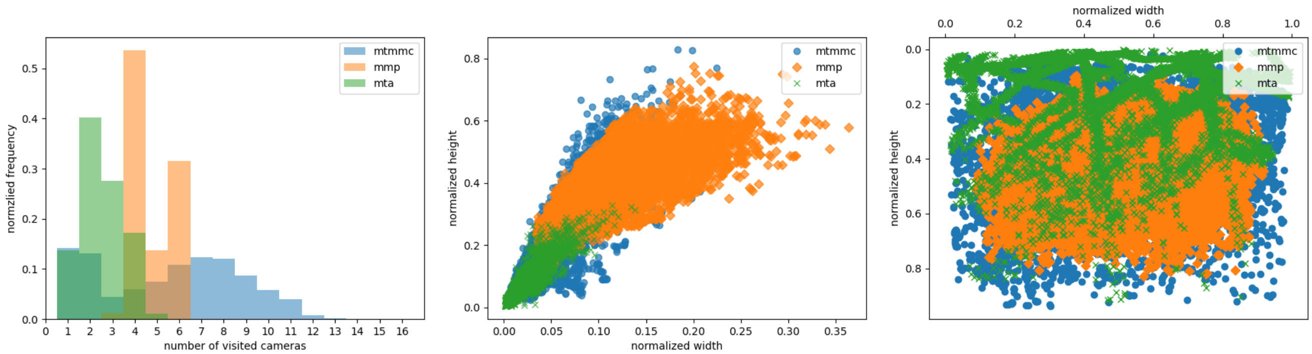

Figure 2. **MTMC Dataset statistics comparison.** We compare **MTMMC** dataset with the current largest **simulated MTA** [27] and **real world MMPTrack** [20] datasets. Each visual summarizes specific statistics of each dataset : (a) The instance tracks by the number of visited cameras, (b) The joint distribution of instance normalized width and height, (c) The instance track centers plotted over normalized image coordinates.

3.2. Data Collection

The data collection was conducted over two days, capturing different seasons for each environment: summer for the factory and fall for the school campus. To ensure a high degree of accuracy in temporal alignment, we employed a precise global time-stamping method for space-time synchronization. For potential frame drops, we meticulously inspected the video sequences and made adjustments by aligning timestamps and interpolating missing data. To prevent any privacy issues, we recruited 623 actors of varying ages and genders and obtained data release agreements. We ensured that all participating actors were compensated for their time and efforts. Furthermore, we conducted de-identification for the 107 non-actors involved in the recordings.

Each video last in five-and-a-half-minutes per scenario, with 12 scenarios from the factory and 13 scenarios from the school campus. We allowed the actors to improvise their actions, provided they fit the given circumstances. For instance, actors could move luggage in the factory or play ball at school, resulting in a wide variety of behaviors being captured. Moreover, we instructed the actors to change their clothes for each scenario to ensure diverse appearances.

Notably, our new MTMMC dataset significantly improves upon the Duke-MTMC dataset [46], which was collected within a narrow 1.5-hour window on a single day on campus. By extending the breadth and diversity of our collection process, we aim to provide a more solid foundation for the development of robust tracking systems.

3.3. Data Annotation

We designed an annotation pipeline to separate the single-camera tracking and the multi-camera association tasks. The single-camera tracking involves generating bounding boxes of person tracks within a camera, while the multi-camera association involves assigning consistent person-IDs across multiple cameras. By dividing the tasks, we can assign inexperienced annotators to the former and skilled workers to the latter. The reviewers carefully check the quality of the completed labels from the annotators, and this process is repeated several times until no critical errors are visible. More details are in the appendix.

Single Camera Tracking. To annotate the set of 400 videos (16 cameras $\times$ 25 scenarios), we tasked annotators with drawing bounding boxes and assigning track-IDs to each person in the video. We collected the annotations in a semi-automatic manner, as described in previous works [54, 57]. First, the annotators tracked and labeled the person in the keyframes, which were selected every five frames in a video. We then used the deepSORT [59] algorithm to generate pseudo tracking boxes by interpolating the annotations between keyframes efficiently. The predicted tracking boxes were then carefully corrected by the annotators. Additionally, to protect the privacy of non-actors, we applied a de-identification process, which involved blurring their faces while remaining the ground truth annotations intact. This process ensured confidentiality of personal information, while simultaneously preserving the data integrity.

Multiple Camera Association. In the next step, we asked annotators to assign consistent track-IDs for the same person *across the cameras* for each scenario. We observed that the semi-automatic labeling approach was not sufficient to achieve satisfactory label quality for this task. Hence, we relied on careful manual labeling. After the initial labeling was completed by the annotators, the reviewers collected person-ID errors using two critical camera constraints. Firstly, one person cannot appear in multiple tracks of the same camera simultaneously. Secondly, one person cannot be visible in the view of two non-overlapping cameras simultaneously. The reviewers also checked for other remaining errors. All the collected errors were then passed to the annotators, who corrected them. The refining process was iterated twice to guarantee high-quality labels.

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

Table 2. **Detection Results.**

3.4. Data Splits

We split the MTMMC dataset into three subsets. The **train** set includes 14 scenarios (7 from the factory and 7 from the campus), the **validation** set includes 5 scenarios (3 from the factory and 2 from the campus), and the **test** set includes 6 scenarios (2 from the factory and 4 from the campus).

4. Experiments

We present various experimental setups and benchmark their performance using the new MTMMC dataset. For the evaluation, we use standard metrics such as mAP for *detection*, Rank1 and mAP for *re-identification*, and CLEAR MOT and IDF1 for *tracking*. The experiments are conducted on the **train** and **validation** sets of the dataset. Detailed setup specifications are in the appendix.

4.1. Sub Tasks: Detection, Re-ID, and MOT

Person Detection We evaluate the efficacy of our dataset in training person detectors for tracking applications. We utilized two well-established detectors, Faster RCNN [44] and YOLOX [19], and investigate how well these models generalize and perform when trained on task-specific versus generic datasets. Specifically, we trained models MTMMC-Person and COCO-Person datasets and then tested their generalization performance using the MOT17 [39] dataset, which presents a variety of real-world tracking scenarios. The COCO-Person is a subset of the larger COCO [32] dataset and includes 65K natural images that depict humans. To compare fairly, we matched the size of our MTMMC-Person dataset, compiling 60K images sampled at a frame rate of 23 fps from the original video footage.

As shown in Table. 2, models trained on the MTMMC-Person dataset consistently outperformed those trained on COCO-Person during the MOT17 evaluations. This suggests that the specificity of the training data to the end-use scenario is crucial. By design, the MTMMC dataset is tailored to tracking, highlighting diverse human activities, frequent occlusions, varied interactions and non-central camera angles, which are typical in real-world tracking situations. These results validate the importance of contextual alignment between training data and its target application, emphasizing the value of our specialized dataset, MTMMC, for tracking and surveillance applications.

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

Table 3. **Re-Identification Results.**

Person Re-Identification In line with the standard protocols for re-identification (Re-ID) data construction, as outlined in [77, 79], we derived our MTMMC-reID dataset from the larger MTMMC dataset. For our experiments, we used the AGW model [66] as the benchmark.

Re-ID tasks require the identification of individuals across multiple camera views and at different times. Training data characteristics significantly influence the performance of Re-ID systems. The MTMMC-reID dataset, in particular, provides a challenging training environment, as evidenced by the lower Rank-1 accuracy and mAP scores—76.0 and 45.6, respectively—compared to other datasets (see Table. 3, rows 2-4). These figures highlight the demanding nature of the tracking scenarios within MTMMC-reID.

However, the dataset's complexity is beneficial for model generalization. For instance, when a model trained on the MSMT17 [58] dataset is evaluated on Market-1501 [77], performance drops (to 64.3 Rank-1 and 34.2 mAP), indicating a loss of generalizability. Yet, if the same model is trained on MTMMC-reID and tested on Market-1501, it demonstrates better robustness with higher Rank-1 accuracy and mAP scores (66.5 and 35.4, respectively) compared to the MSMT17 training (refer to Table. 3, rows 5-6). These results imply that despite the intrinsic challenges of MTMMC-reID, models trained on it are better equipped to handle new, unseen environments, underscoring the value of rigorous training environments for improved real-world applicability.

Multi Object Tracking Multi-object tracking (MOT) is a task that requires the detection and tracking of multiple objects, often people, through a sequence of video frames. The challenge lies in keeping consistent object identities despite movement, occlusions, and environmental changes. In our experiment, we employed four state-of-the-art trackers: JDE [56], QDTrack [40], CenterTrack [82], and ByteTrack [74], and our analysis focuses on three main aspects:

1. **Training and Evaluating on the Same Dataset**: When models are both trained and evaluated on the same dataset, they exhibit lower performance on the MTMMC compared to the MOT17 dataset. For instance, JDE, achieved an IDF1 score of 42.4% on MTMMC, whereas the same model yielded an improved IDF1 of 63.6% on MOT17. This trend is consistent across all tested models, indicating that MTMMC presents a more challenging testbed.

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

Table 4. **Multi Object Tracking Results.** Following the previous works, we use additional person detection data: CH denotes CrowdHuman [47], MIX indicates combined datasets of Caltech Pedestrian [12], Citypersons [73], CUHK-SYS [63], PRW [78] and ETH [14].

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

Table 5. **Pre-Training** on MTMMC and MOTSynth.

2. **Training and Evaluating on Different Datasets**: When training and evaluation datasets differed, we observed a pattern where models trained on MTMMC generally outperformed those trained on MOT17 when evaluated on the alternate dataset. For example, ByteTrack, after being trained on MTMMC and tested on MOT17, reached an IDF1 score of 69.1% and MOTA of 55.9%, which is closer to the practical upper bounds observed when trained and tested on MOT17 (IDF1 of 76.8% and MOTA of 75.0%). In contrast, when ByteTrack was trained on MOT17 and evaluated on MTMMC, it achieved a much lower IDF1 of 40.2% and MOTA of 56.8%, versus its upper-bound performance on MTMMC (IDF1 of 64.8% and MOTA of 89.7%). This suggests that the complex and diverse tracking environments found in MTMMC contribute to the development of more robust and generalizable model features.

   Notably, the above two trends within multi-object tracking mirror the tendencies observed in our Re-ID experiments. This consistency reinforces the notion that training on more complex and diverse environments effectively enhances the models' ability to generalize and maintain accuracy when introduced to new domains.

3. **Training on Combined Datasets**: The most compelling results were observed when models were trained on a mixture of both MTMMC and MOT17 datasets. This combined training approach produced the best results on both MTMMC and MOT17 evaluations. It implies that the MTMMC provides a complementary training signal to the MOT17. When combined, the diversity and complexity of MTMMC complement the MOT17, leading to a robust tracking model.

In conclusion, these experiments underline the importance of dataset diversity and complexity in training multi-object tracking models. The demanding context provided by MTMMC help to forge models that can handle real-world complexities effectively.

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

In this study, we evaluate the efficacy of real-world data in improving MOT models by employing our MTMMC dataset as a foundational training set. We utilized the QDTrack [40] as our base tracker and conducted experiments to measure its performance on the MOT17 benchmark. These experiments

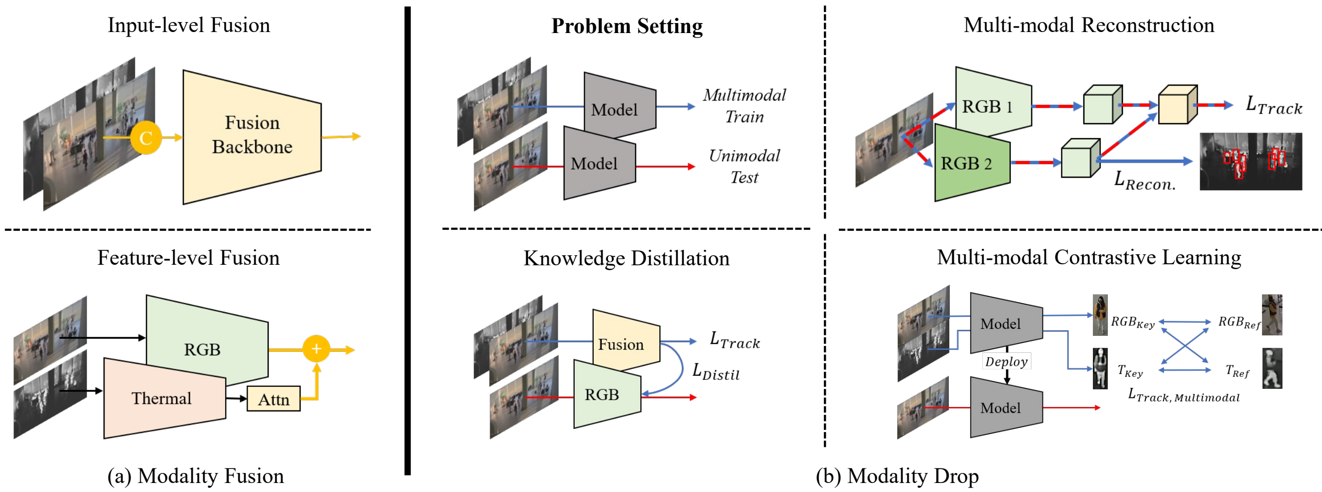

Figure 3. **Multi-modal Learning Setups and Baselines.** (a) presents the concept of modality fusion with both input-level and feature-level fusion techniques integrating thermal data with RGB for enhanced object tracking. (b) outlines the modality drop scenario, where the model trained on combined RGB and thermal data is tested solely on RGB data, using methods like multi-modal reconstruction, knowledge distillation, and multi-modal contrastive learning.

involved pre-training the model on the MTMMC dataset and subsequently fine-tuning it on MOT17. Additionally, we drew comparisons with models pre-trained on the MOTSynth dataset [15], which is a large-scale synthetic dataset derived from extensive simulation within a gaming environment.

As detailed in Table 5, our findings illustrate that the MTMMC dataset, albeit comprising half the number of annotations compared to MOTSynth (0.5M vs. 1M), and without the aid of complex data simulation techniques, still substantially contributes to the tracking accuracy. Notably, models pre-trained on MTMMC yield a MOTA score of 55.3 without fine-tuning (54.1 when pre-trained on MOTSynth) and see an increase to 68.6 with fine-tuning (70.8 when pre-trained on MOTSynth). While MOTSynth commences at a higher baseline, our real-world data, when combined with MOTSynth, demonstrates a remarkable synergy, resulting in a superior IDF1 score of 72.0 post fine-tuning.

These observations underscore the continued relevance of real-world datasets. While the scalability and control offered by synthetic data are appealing, the inherent complexities and variability present in real-world data are crucial for models to learn effectively. The MTMMC dataset, therefore, remains an invaluable resource for achieving high-fidelity tracking performance, and its integration with synthetic data further enhances this advancement.

4.3. Multi-modal Learning: Setups and Baselines

Multi-modal learning aims to improve scene understanding by leveraging complementary information from different sensor modalities. In this context, we explore how thermal data, when paired with RGB data, can enhance object tracking. This question stems from existing literature that demonstrates the benefits of such combinations in other domains [2, 53, 81]. Our research extends these concepts into tracking scenarios using QDTrack [40] as the base tracker. We present two new learning setups, modality fusion and drop, illustrated in Fig. 3-(a) and (b), respectively. We provide more detailed setup specifications and additional analyses in the appendix. Here, we briefly introduce the high-level concepts of the setups and then discuss the key results.

Modality Fusion We begin with modality fusion, focusing on the explicit integration of thermal data into RGB-based tracking models. This involves comparing both *input* and *feature-level* fusion methods against RGB and thermal-only baselines. We evaluate the benefits of thermal data incorporation, when it is directly available for both train and test.

Modality Drop The modality drop setup presents a more challenging scenario. Here, the model is trained on both RGB and thermal data but is evaluated solely on RGB data. The rationale is that, during training, the model can learn generalized feature representations that are robust even when a modality is absent during testing. We introduce three strategies to harness RGB-T data effectively during training: *knowledge distillation*, *multi-modal reconstruction*, and *multi-modal contrastive learning*.

One practical application is using a multimodally trained tracking model in an unimodal tracking system. For instance, consider CCTV surveillance systems, which predominantly rely on RGB cameras often due to hardware or budget constraints. Our goal is to train the model using datasets like MTMMC, which contain both RGB and thermal data, and then test its effectiveness in environments that only provide RGB data. Essentially, we aim to determine if the model can learn generic features from the combined RGB and thermal data during training, and preserve its tracking capabilities in the absence of thermal data during testing.

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

Table 6. **Multi-modal learning results.**

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

Table 7. **Multi-Target Multi-Camera Tracking Results** in MTMMC. For the efficient evaluation, we temporally sub-sampled the videos in 1FPS. H. Cluster denotes hierarchical clustering. The averaged results of all the testing scenarios are shown.

Results The results in Table. 6-(a) showcase the performance gains from modality fusion. The integration of thermal data at both the input (RGBT-I) and feature level (RGBT-F) with the base RGB data results in improved performance, compared to using RGB or thermal data in isolation. Notably, the RGBT-F approach, achieves the highest overall performance, with an IDF1 score of 53.9 and MOTA of 86.0. This suggests that thermal data, when fused at the feature level, provides a more discriminative tracking representation.

In Table. 6-(b), we summarize the performance in the modality drop setup. We simulate the modality drop scenario, by training the model using both RGB and thermal data in the MTMMC, and evaluate or optionally fine-tune the model using MOT17, which only provides RGB data. Here, the 'without fine-tuning' demonstrates how well the features learned from the combined multimodal data (RGB+T) transfer directly to the RGB domain. On the other hand, 'with fine-tuning' evaluates how effectively these learned features serve as initialization for further refinement. Without fine-tuning, Knowledge Distillation (KD) lags in performance (IDF1: 55.1, MOTA: 43.2), which is likely due to its strong dependence on thermal data imposed during distillation, resulting in a weaker generalization ability. In contrast, the Multi-modal Contrastive method shows a relatively high resilience (IDF1: 59.7, MOTA: 48.4), suggesting it learns modality-invariant features through contrastive learning, which confers strong generalization. With fine-tuning, KD exhibits a marked improvement (IDF1: 70.5, MOTA: 68.0), indicating its potential once adapted to the RGB domain. Conversely, the Multi-modal Contrastive method sees only a marginal increase after fine-tuning (IDF1: 68.3, MOTA: 67.3). It is important to note that generalizable features do not necessarily equate to an optimal initialization for RGB-specific fine-tuning. We recognize the further investigations are necessary to fully understand the underlying mechanisms, and we leave this for future studies.

4.4. Multi-modal MTMC

Multi-target multi-camera (MTMC) expands upon MOT by requiring the identification of multiple targets across various camera views. We build a strong baseline model to benchmark the MTMC scores on our new MTMMC dataset. Specifically, we integrate the multi-object tracker and person Re-ID networks, QDtrack [40] and BoT [36] to generate the tracklet-level feature representation. Upon this, we study the performance of two leading multi-camera association (MCA) methods, optimization-based [21] and clustering-based [27].

In Table. 7-(a), the results show that the hierarchical clustering-based MCA method [27] outperformed the optimization-based approach [21], which required heavy hyper-parameter tuning. Table 7-(b) presents the results following the integration of thermal information on the clustering-based method. The feature-fusion approach again resulted in more accurate multi-camera tracking. As a dataset and evaluation paper, we focus on establishing baseline models and benchmark scores to set a stage for followup researches. We hope to see numerous advanced multi-modal tracking models presented upon our results.

5. Conclusion

We have presented the MTMMC dataset—a large-scale, real-world, multi-modal tracking benchmark designed to advance MTMC tracking. Through our extensive experiments, we have demonstrated its efficacy in improving the performance of various sub-tasks and have highlighted its synergistic use with synthetic data for pre-training. Additionally, we introduced two new multi-modal learning setups—modality fusion and drop—and developed robust baseline models for multi-modal MTMC tracking. We hope that our contributions will reinvigorate research in MTMC and will spark new innovations in multi-modal tracking technologies, ushering in a new era of intelligent tracking systems.

Acknowledgement This work was partially supported by the NRF (NRF-2020M3H8A1115028, FY2021).

References

[1] Mahdi Abavisani, Hamid Reza Vaezi Joze, and Vishal M Patel. Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1165–1174, 2019. 2

[2] Ganbayar Batchuluun, Dat Tien Nguyen, Tuyen Danh Pham, Chanhum Park, and Kang Ryoung Park. Action recognition from thermal videos. *IEEE Access*, 7:103893–103917, 2019. 7

[3] Jerome Berclaz, Francois Fleuret, Engin Turetken, and Pascal Fua. Multiple object tracking using k-shortest paths optimization. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1806–1819, 2011. 2, 3

[4] Francesco Bonchi, David Garcia-Soriano, and Edo Liberty. Correlation clustering: from theory to practice. In *KDD*, page 1972, 2014. 3

[5] Michael Bredereck, Xiaoyan Jiang, Marco Körner, and Joachim Denzler. Data association for multi-object tracking-by-detection in multi-camera networks. In *2012 Sixth International Conference on Distributed Smart Cameras (ICDSC)*, pages 1–6. IEEE, 2012. 3

[6] Lijun Cao, Weihua Chen, Xiaotang Chen, Shuai Zheng, and Kaiqi Huang. An equalised global graphical model-based approach for multi-camera object tracking. *arXiv preprint arXiv:1502.03532*, 8, 2015. 2, 3

[7] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. The wildtrack multi-camera person dataset. *arXiv preprint arXiv:1707.09299*, 2017. 2, 3

[8] Long Chen, Haizhou Ai, Rui Chen, Zijie Zhuang, and Shuang Liu. Cross-view tracking for multi-human 3d pose estimation at over 100 fps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3279–3288, 2020. 3

[9] Abir Das, Anirban Chakraborty, and Amit K Roy-Chowdhury. Consistent re-identification in a camera network. In *European conference on computer vision*, pages 330–345. Springer, 2014. 3

[10] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision*, pages 436–454. Springer, 2020. 2

[11] Christophe De Vleeschouwer, Fan Chen, Damien Delannay, Christophe Parisot, Christophe Chaudy, Eric Martrou, Andrea Cavallaro, et al. Distributed video acquisition and annotation for sport-event summarization. *NEM summit*, 8, 2008. 2

[12] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311, 2009. 6

[13] Tiziana D'Orazio, Marco Leo, Nicola Mosca, Paolo Spagnolo, and Pier Luigi Mazzeo. A semi-automatic system for ground truth generation of soccer video sequences. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 559–564. IEEE, 2009. 2

[14] Andreas Ess, Bastian Leibe, Konrad Schindler, and Luc Van Gool. A mobile vision system for robust multi-person tracking. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 6

[15] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10849–10859, 2021. 7

[16] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *Proceedings of the IEEE international conference on computer vision*, pages 3038–3046, 2017. 2

[17] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–6. IEEE, 2009. 2, 3

[18] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):267–282, 2007. 2

[19] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 5

[20] Xiaotian Han, Quanzeng You, Chunyu Wang, Zhizheng Zhang, Peng Chu, Houdong Hu, Jiang Wang, and Zicheng Liu. Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark. *arXiv preprint arXiv:2111.15157*, 2021. 2, 3, 4

[21] Yuhang He, Xing Wei, Xiaopeng Hong, Weiwei Shi, and Yihong Gong. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Transactions on Image Processing*, 29:5191–5205, 2020. 3, 8

[22] Yunzhong Hou, Zhongdao Wang, Shengjin Wang, and Liang Zheng. Adaptive affinity for associations in multi-target multi-camera tracking. *IEEE Transactions on Image Processing*, 31:612–622, 2021. 3

[23] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krahenbuhl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5390–5399, 2019. 2

[24] Na Jiang, SiChen Bai, Yue Xu, Chang Xing, Zhong Zhou, and Wei Wu. Online inter-camera trajectory association exploiting person re-identification and camera topology. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1457–1465, 2018. 3

[25] Ugur Kart, Joni-Kristian Kamarainen, and Jiri Matas. How to make an rgbd tracker? In *proceedings of the european conference on computer vision (ECCV) Workshops*, pages 0–0, 2018. 2

[26] Ugur Kart, Alan Lukezic, Matej Kristan, Joni-Kristian Kamarainen, and Jiri Matas. Object tracking by reconstruction with view-specific discriminative correlation filters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1339–1348, 2019. 2

[27] Philipp Kohl, Andreas Specker, Arne Schumann, and Jurgen Beyerer. The mta dataset for multi-target multi-camera pedestrian tracking by weighted distance aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1042–1043, 2020. 2, 3, 4, 8

[28] Matej Kristan, Aleš Leonardis, Jiří Matas, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Martin Danelljan, Luka Čehovin Zajc, Alan Lukežič, Ondrej Drbohlav, et al. The eighth visual object tracking vot2020 challenge results. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 547–601. Springer, 2020. 2

[29] Cheng-Hao Kuo, Chang Huang, and Ram Nevatia. Inter-camera association of multi-target tracks by on-line learned appearance affinity models. In *European conference on computer vision*, pages 383–396. Springer, 2010. 2, 3

[30] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 33–40, 2016. 2

[31] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019. 2

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[33] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[35] Ye Liu, Xiao-Yuan Jing, Jianhui Nie, Hao Gao, Jun Liu, and Guo-Ping Jiang. Context-aware three-dimensional mean-shift with occlusion handling for robust object tracking in rgb-d videos. *IEEE Transactions on Multimedia*, 21(3):664–677, 2018. 2

[36] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 8

[37] Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 166–183, 2018. 2

[38] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. *arXiv preprint arXiv:2101.02702*, 2021. 2

[39] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 2, 5

[40] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021. 2, 5, 6, 7, 8

[41] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Per-clip video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1352–1361, 2022. 2

[42] Kha Gia Quach, Pha Nguyen, Huu Le, Thanh-Dat Truong, Chi Nhan Duong, Minh-Triet Tran, and Khoa Luu. Dyglip: A dynamic graph model with link prediction for accurate multi-camera multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13784–13793, 2021. 3

[43] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 2

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 5

[45] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6036–6046, 2018. 3

[46] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 2, 3, 4

[47] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6

[48] Ukcheol Shin, Kwanyong Park, Byeong-Uk Lee, Kyunghyun Lee, and In So Kweon. Self-supervised monocular depth estimation from thermal images via adversarial multi-spectral adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5798–5807, 2023. 2

[49] KA Shiva Kumar, KR Ramakrishnan, and GN Rathna. Distributed person of interest tracking in camera networks. In *Proceedings of the 11th International Conference on Distributed Smart Cameras*, pages 131–137, 2017. 3

[50] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2

[51] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[52] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):104–119, 2019. 2

[53] Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019. 2, 7

[54] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10776–10785, 2021. 4

[55] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, pages 107–122. Springer, 2020. 2

[56] Zhongdao Wang, Liang Zheng, Yixuan Liu, and Shengjin Wang. Towards real-time multi-object tracking. *The European Conference on Computer Vision (ECCV)*, 2020. 5

[57] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, et al. Step: Segmenting and tracking every pixel. *arXiv preprint arXiv:2102.11859*, 2021. 4

[58] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 5

[59] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 2, 4

[60] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Bridging images and videos: A simple learning framework for large vocabulary video object detection. In *European Conference on Computer Vision*, pages 238–258. Springer, 2022. 2

[61] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Tracking by associating clips. In *European Conference on Computer Vision*, pages 129–145. Springer, 2022. 2

[62] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12352–12361, 2021. 2

[63] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017. 6

[64] Dan Xu, Wanli Ouyang, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning cross-modal deep representations for robust pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5371, 2017. 2

[65] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10725–10733, 2021. 2

[66] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 5

[67] Quanzeng You and Hao Jiang. Real-time 3d deep multi-camera tracking. *arXiv preprint arXiv:2003.11753*, 2020. 3

[68] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 2

[69] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 2

[70] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain. Cross-modality interactive attention network for multispectral pedestrian detection. *Information Fusion*, 50:20–29, 2019. 2

[71] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5127–5137, 2019. 2

[72] Shu Zhang, Elliot Staudt, Tim Faltemier, and Amit K Roy-Chowdhury. A camera network tracking (camnet) dataset and performance baseline. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 365–372. IEEE, 2015. 2, 3

[73] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 6

[74] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. 5

[75] Zhimeng Zhang, Jianan Wu, Xuan Zhang, and Chi Zhang. Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. *arXiv preprint arXiv:1712.09531*, 2017. 3

[76] Long Zhao, Xi Peng, Yuxiao Chen, Mubbasir Kapadia, and Dimitris N Metaxas. Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6528–6537, 2020. 2

[77] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 5

[78] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1367–1376, 2017. 6

[79] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. 5

[80] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Reranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1318–1327, 2017. 3

[81] Wujie Zhou, Jinfu Liu, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Gmnet: graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation. *IEEE Transactions on Image Processing*, 30:7790–7802, 2021. 2, 7

[82] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 2, 5