Supplementary Material:
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

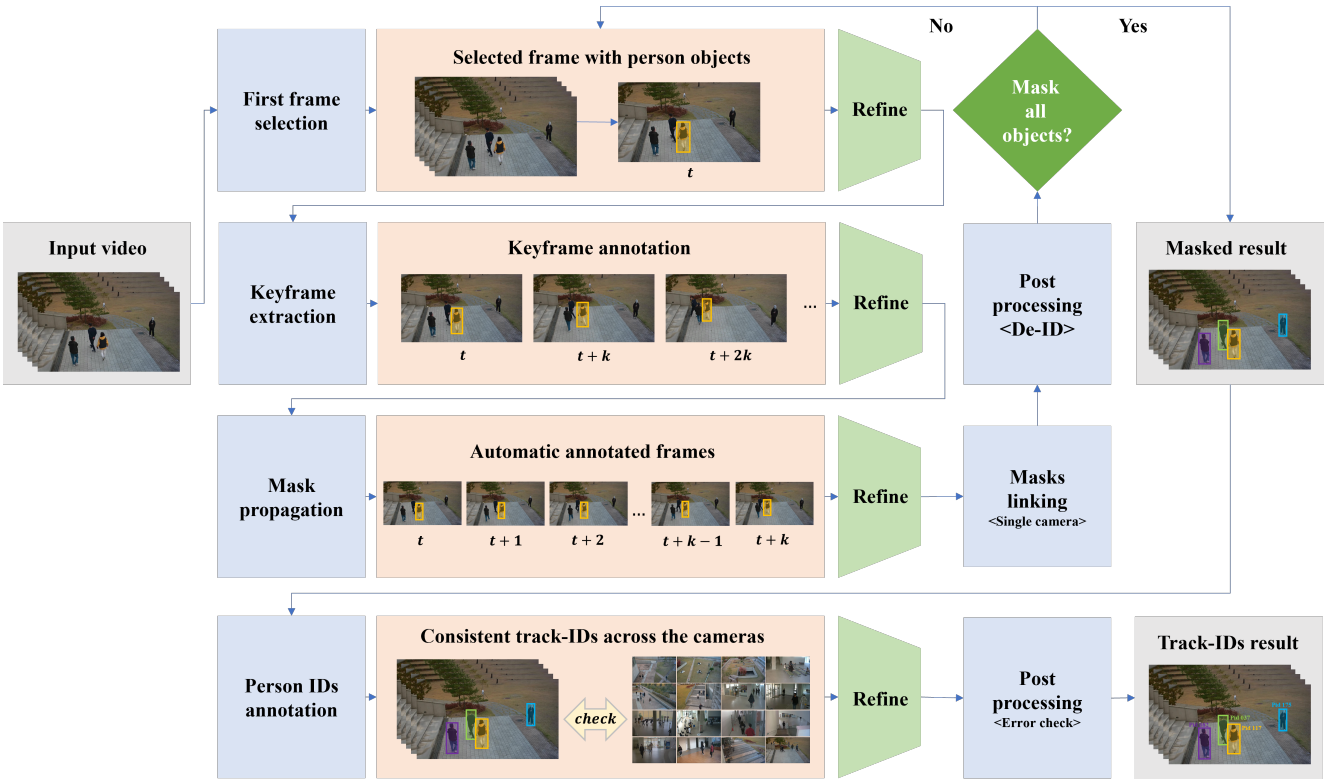


Figure 1. **Single-camera tracking annotation pipeline.** We adopt the semi-automatic labeling approach. The workers first label the key frames and then the annotations for the other frames are interpolated based on the model predictions.

In this supplementary material, we present detailed information on the following aspects:

A) Details of Annotation,
B) Experimental Setup Specifications,
C) Supplementary Experiments,
D) Specifications of Camera Hardware,
E) Overview of the MTMMC Dataset,
F) Licenses of the Datasets Used,
G) Video Demonstration, and
H) Discussion of Ethical Considerations.

1. Annotation Details

1.1. Annotation Pipeline

Our annotation pipeline is illustrated in Fig. 1. We separate the single-camera tracking task from the multi-camera association. The annotation tool is built upon the CVAT [1], an open-source vision annotation tool. We further enabled several functionalities such as uploading large-scale videos, efficient task management of crowd workers, and text description translations.

[1]https://github.com/maheriya/cvat

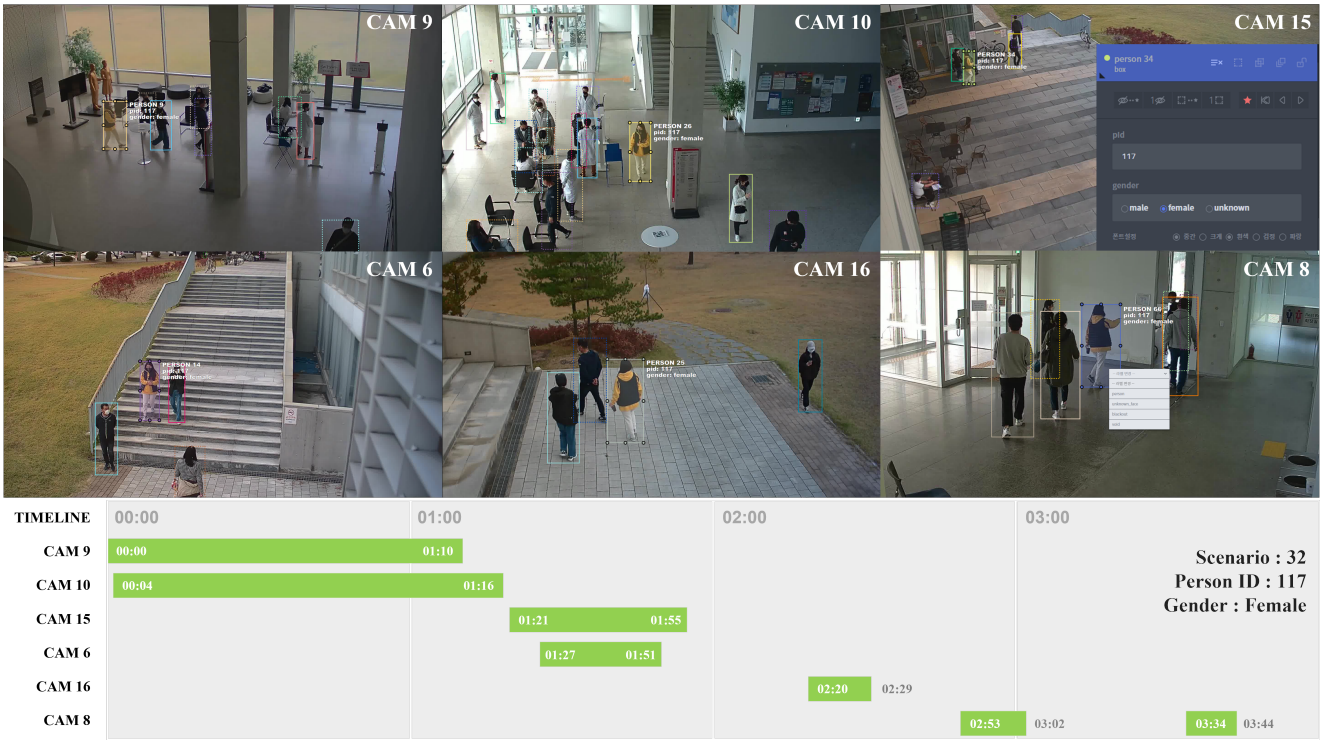
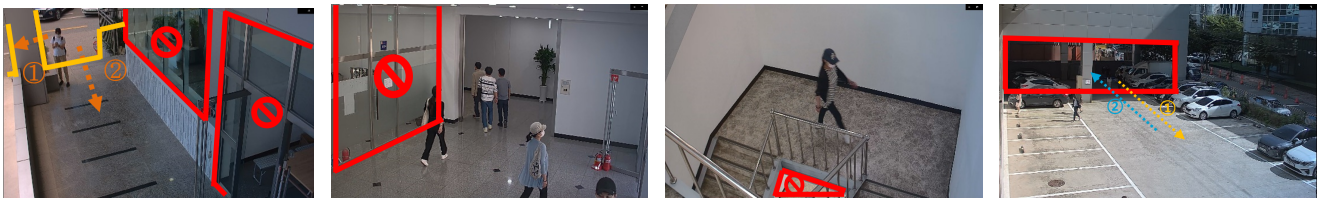


Figure 2. **Multi-camera association.** The workers are instructed to assign consistent PIDs for the same person across the cameras for each scenario.

(a) (b) (c) (d) (e)

Figure 3. **When to assign dummy person IDs.** (a),(b) horizontal and vertical truncation (c),(d) severe lighting (e) small-scale.



(a) (b) (c) (d)

Figure 4. **When to ignore the labeling.** (a),(b) reflection on transparent objects (c) severe truncation (d) poor lighting.

(a) (b) (c) (d)

Figure 5. **Amodal annotation example.**

Single-camera tracking We employed 189 workers, 12 reviewers, and 2 project managers for our annotation task. The project managers provided annotation guidelines and managed the scheduling. As detailed in our main paper, we adopted a semi-automatic labeling approach. We manually annotated keyframes (with a default temporal stride of 5 frames) and used the deepSORT algorithm [22] for interpolating annotations on intermediate frames. Workers had the flexibility to adjust the temporal stride. To ensure high-quality annotations, we conducted rigorous peer reviews. In instances where bystanders were captured, we applied face blurring in post-processing to address privacy concerns. On average, each worker annotated 525 images per day. All annotations were saved in JSON format.

Multi-camera Association For this task, we selected a team of 8 highly skilled workers, supported by 8 reviewers and 2 project managers. Initially, we attempted model prediction-based labeling, but it failed to meet our internal quality standards. Consequently, as illustrated in Fig. 2, we shifted to manual labeling. Workers used the labels generated from the single-camera tracking phase as a base, checking for errors and matching labels across different cameras to assign final Person IDs (PIDs). In situations with ambiguous identifications, workers were permitted to use a placeholder label of '0'. To guarantee the accuracy of our data, we enforced two essential camera constraints and iteratively corrected any discrepancies until we observed no significant errors.

1.2. Annotation Instructions

To ensure high-quality annotations, we established specific guidelines for workers to handle exceptional cases. These guidelines were designed to maintain consistency and accuracy in challenging annotation scenarios.

- **Dummy Person ID** In instances where individuals were challenging to identify due to significant changes in appearance, scale, or lighting, workers assigned a placeholder ID of '0'. This practice was crucial for maintaining data integrity in cases where person identification was ambiguous or unreliable (refer to Fig. 3).
- **Ignore Area** Our guidelines specified that reflections of persons on transparent surfaces, such as glass doors, should be excluded from annotations. Similarly, extreme cases of body part truncation or poor lighting conditions were treated as grey areas and omitted from the dataset to ensure annotation quality (refer to Fig. 4).
- **Amodal Annotation** Following the MOT17 annotation standard [16], we instructed workers to label objects amodally. This approach involved marking occlusions distinctly from visible parts using dotted lines, enhancing the precision of our dataset (refer to Fig. 5).

2. Experimental Details

2.1. Datasets

We provide the details of the additional data source used for the experiments in the main paper.

COCO2017 COCO [13] includes object detection and segmentation labels with 118k training images and 5k validation images. We collected only the person detection labels, COCO-Person, and used them for the person detection experiments.

Market-1501 Market-1501 [26] is a popular person re-ID dataset collected from six outdoor cameras. It contains 32,668 bounding boxes of 1,501 identities. To simulate the realistic scenarios, Deformable Part Model (DPM) [7] is employed to produce bounding boxes of pedestrians. We used it for a person re-identification experiments.

MSMT17 MSMT17 [21] is a challenging person re-identification dataset with 126,441 bounding boxes of 4,101 identities captured by 15 outdoor cameras and predicted using Faster RCNN [19]. The dataset includes complex scenes and significant variations in lighting and viewpoints to provide challenging cases. We utilized MSMT17 for our person re-identification experiments.

MOT17 MOT17 [16] is a widely used multi-object tracking benchmark that includes 7 training and 7 testing videos with challenging scenarios, such as frequent occlusions and crowd scenes. We split each training sequence into two halves and used the first half-frames for training and the second half for validation. We utilized MOT17 for single-camera tracking experiments.

MOTSynth MOTSynth [6] is a large-scale synthetic dataset created using the photorealistic video game Grand Theft Auto V. It is designed for person tracking and segmentation in urban scenarios and contains 764 full-HD videos, 1.3M frames, and 33M person instances. We utilized the official train and validation split and applied it to our transfer learning experiments.

2.2. Implementation Details

Training Our framework is constructed using two well-established public codebases: *mmtracking* [3] and *fast-reid* [8]. We adhere closely to the default training recipes provided by these codebases, including data augmentation techniques and training schedules detailed in Table. 1. For implementation, we utilized Pytorch v1.10 and CUDA v11.3, executing our models on a robust hardware setup equipped with an AMD EPYC 7352 (2.3GHz) CPU and an NVIDIA RTX A6000 GPU, ensuring efficient processing and analysis.

Sub-task	Model	epoch	fine-tune	optimizer	lr	momentum	weight_decay
Detection	Faster R-CNN	24	4	SGD	0.02	0.9	0.0001
Re-ID	AGW	120	-	adam	0.00035	-	0.0005
	BOT	120	-	adam	0.00035	-	0.0005
MOT	JDE	30	30	SGD	0.01	0.9	0.0001
	QDTrack	4	4	SGD	0.02	0.9	0.0001
	CenterTrack	70	best	adam	0.0125	-	0.0001
	ByteTrack	300	40	SGD	0.01	0.9	0.0005

Table 1. **Subtasks Training Recipes**

Method	In.	Attn.	Asy.	IDF1	MOTA	mAP
RGB	-	-	-	53.0	84.5	92.8
(1)	Add.	BAM	✓	53.7	85.3	93.1
(2)	Diff.	Cha.	✓	53.5	85.1	93.0
	Diff.	Spa.	✓	53.7	85.4	93.3
(3)	Diff.	BAM	✗	53.8	85.6	93.2
RGBT-F	Diff.	BAM	✓	**53.9**	**86.0**	**93.5**

(a) Feature-level Fusion

Modality	Network	# contra.loss pair	IDF1	MOTA	mAP
RGB	baseline	1	55.7	43.8	55.6
RGB-T	parallel	2	56.1	44.5	57.5
		4	56.8	44.3	57.6
	share	2	55.6	43.5	56.7
		4	**59.7**	**48.4**	**59.7**
		6	57.3	44.5	57.3
		8	57.5	45.4	58.1

(b) Multi-modal contrastive-learning

Table 2. **Ablation studies of multi-modal learning models on MTMMC.**

Multi-model Learning In this section, we detail the design of two proposed multi-modal learning setups.

1) Modality Fusion

- **Input-level fusion.** We employ channel-wise concatenation [5] to combine RGB and Thermal inputs into a 4-channel RGB-T input. This involves modifying the first convolutional layer in the backbone to accept 4 channels instead of 3, initializing the additional channel's weight as the average of the RGB channel weights.
- **Feature-level fusion.** Building upon prior research [4, 25, 27], we have developed a conditional attention module to utilize multi-modal data more effectively. This module uses a BAM attention block [18], processing the differential between RGB and thermal features to explicitly highlight the complementary nature of these modalities. The fusion process is executed as $F_{fuse} = F_{rgb} + BAM(F_{thermal} - F_{rgb}) \otimes F_{thermal}$, where F_{fuse} represents the fused feature output and $\otimes$ is element-wise multiplication. We integrate this attention module at various levels within the Feature Pyramid Network (FPN [12]), training it end-to-end without additional adjustments.

To empirically validate our design choices, we conducted ablation studies as shown in Table 2-(a). We investigated three important design choices: (1) input for the attention module, (2) attention design, and (3) asymmetric encoders. Firstly, we found that forwarding the feature difference (Diff.) between the thermal and RGB performed better than their addition (Add.). Secondly, while using either channel (Cha.) or spatial (Spa.) attention resulted in performance improvements over the simple baseline model, jointly using them (BAM) yielded the best performance, as noted in the original paper [18]. Finally, the asymmetric encoder design, which is intended to properly cope with the input information density, resulted in better tracking performance.

2) Modality Drop

- **Knowledge distillation.** To transfer the knowledge learned from the RGBT-F model to the standard RGB model, we distill the final level of FPN features to the RGB model. We implement the distillation loss using the MSE loss function, defined as $L_{Distill} = L_{MSE}(F_{rgb}, F_{fuse})$. The final objective function is defined as $L = L_{Track} + \lambda L_{Distill}$, where L_{Track} represents the total loss to train the base tracker [17]. The λ is set to 0.1.
- **Multi-modal reconstruction.** Motivated by the MSDN framework [23], we incorporate the multi-modal reconstruction loss into the base tracker design. Specifically, we use two identical backbones, B_1 and B_2, and extract RoI features F_1 and F_2 from the RGB input. We then use F_2 to reconstruct the corresponding thermal information with a single deconvolution layer, which enforces F_2 to encode the RGB-to-Thermal correlations explicitly. The fused features of F_1 and F_2 are then forwarded to the tracking head, implicitly encoding the thermal information without its presence at test time. The total loss is $L = L_{track} + \lambda L_{recon}$, where λ is set to 1.0.

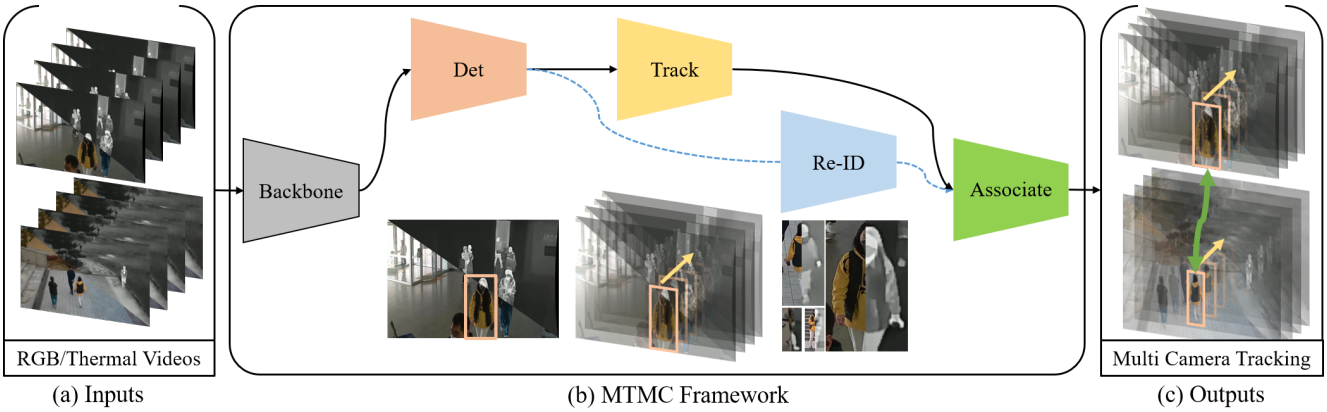


Figure 6. **Overview of multi-target multi-camera tracker.**

- **Multi-modal contrastive-learning.** To enhance the multi-view contrastive learning, we made two adaptations to the multiple positive contrastive loss [17] that match the RGB instance features across the key and reference frames. Firstly, we allowed the instance feature matching within the frames, such as (key-key). Secondly, we included the instance thermal features, resulting in new feature combinations, such as RGB-T or T-T. The anchor features were selected from the key frame, and the positive and negative features were selected from the reference frame. The following combinations are investigated:
  - $L_{contra}^{2-pair} = RGB_{Key} - RGB_{Ref} + T_{Key} - T_{Ref}$
  - $L_{contra}^{4-pair} = L_{contra}^{2-pair} + RGB_{Key} - T_{Ref} + T_{Key} - RGB_{Ref}$
  - $L_{contra}^{6-pair} = L_{contra}^{4-pair} + RGB_{Key} - T_{Key} + T_{Key} - RGB_{Key}$
  - $L_{contra}^{8-pair} = L_{contra}^{6-pair} + RGB_{Ref} - T_{Ref} + T_{Ref} - RGB_{Ref}$

  In Table. 2-(b), we investigate the impact of multi-view contrastive learning along with the model design. Firstly, we find that multi-view matching generally outperforms the original RGB-based single-view matching. Secondly, sharing the backbone encoder for two different input sources leads to better results.

In our empirical evaluation, we demonstrated that our Modality Fusion method significantly enhances tracking performance. Additionally, the Modality Drop approach effectively encodes multi-modal correlations, improving tracker generalizability. Building on these findings, we designed a multi-modal Multi-Target Multi-Camera (MTMC) system, as detailed in the architectural overview in Fig. 6. Our system employs a two-stage approach. Initially, it generates tracklets for each video using a tracking-by-detection mechanism. These tracklets are then linked within each video through a Multi-Camera Association (MCA) process. In the main paper, we investigate the two primary MCA approaches: optimization-based [9] and clustering-based methods [11]. Moreover, we have validated the effectiveness of incorporating additional thermal information as a prior, which significantly enhances the tracker's capabilities.

Method	Train	Rank 1	mAP
AGW	Market-1501	98.0	65.2
	MSMT17	94.0	69.9
	MTMMC-reID	**98.0**	**72.5**
BOT	Market-1501	96.0	64.4
	MSMT17	94.0	68.6
	MTMMC-reID	**98.0**	**69.5**

Table 3. **Cross-domain re-identification results.**

Method	Modality	Rank 1	mAP
AGW	RGB	78.8	45.7
	RGBT-I	79.2	47.3
	RGBT-F	**80.7**	**48.4**
BOT	RGB	74.4	42.7
	RGBT-I	75.1	44.1
	RGBT-F	**77.7**	**44.6**

Table 4. **Multi-modal re-identification results.**

Method	Train set		Eval on MTMMC		Eval on MOT17	
	MTMMC	MOT17	IDF1	MOTA	IDF1	MOTA
BoT-SORT	✓		64.7	89.2	70.8	56.7
		✓	40.7	58.5	78.1	75.0
OC-SORT	✓		63.4	88.5	68.8	55.3
		✓	39.2	52.8	74.5	73.5

Table 5. **Multi Object Tracking Results using SOTA trackers.**

3. Additional Experiments

Cross-domain Person Re-Identification To illustrate the broad applicability of our re-ID representation, we conducted cross-domain experiments in Table. 3 using using two Re-ID models, BOT [15] and AGW [24]. Trained on pedestrian datasets (Market-1501 [26], MSMT17 [21], and MTMMC-reID) and tested in sports scenes [20], the model using our dataset demonstrated superior performance, indicating that our dataset provides more generic representations.

Multi-modal Person Re-Identification To further validate the effectiveness of incorporating thermal modality, we conducted multi-modal re-identification experiments. We implemented two fusion approaches: input-level fusion (RGBT-I) and feature-level fusion (RGBT-F). Our findings indicate performance enhancements in both the BOT [15] and AGW [24] models, thereby reaffirming the efficacy of integrating thermal modality in this context.

Multi Object Tracking w. SOTA trackers In Table. 5, we present the results of MOT experiments using two recent models [1, 2]. The results demonstrate that these algorithms exhibit lower performance when trained and tested on our MTMMC dataset compared to their performance on the MOT17 dataset. This indicates that our MTMMC dataset presents a richer array of complexities, emphasizing its significance in training and testing more advanced models.

4. Camera Hardware Specifics

We detail our RGB-Thermal multi-modal camera system, adapted from Hwang et al. [10] for surveillance. It consists of an RGB camera (1920 × 1080 resolution) and a thermal camera (320 × 240 resolution), both with a 50° horizontal field of view and 60 fps frame rate. A hot mirror is included to filter extraneous radiation. For uniformity, thermal images are resized to 1920 × 1080, with border areas discarded during processing. See Fig. 7 (c) and (d) for further details.

5. MTMMC Dataset Details

5.1. Dataset Split

We divided the 25 MTMMC scenarios into three sets, consisting of 14, 5, and 6 scenarios for the training, validation, and testing sets, respectively. The division was based on the meta information to ensure equal and well-distributed representation, shown in Table 6.

5.2. Dataset Statistics

This section presents a detailed analysis of our dataset through three key metrics: (1) objects per frame, (2) tracks per video, and (3) age and gender distribution. These metrics highlight the dataset's unique characteristics, enhancing its effectiveness for multiple object tracking tasks.

Number of Objects per Frame The average number of objects per frame at each site is shown in Figure 9, with error bars indicating standard deviation. Our dataset showcases a significant variation in the number of objects per camera, mirroring real-world environments characterized by challenges like occlusion and diverse movement patterns. A noteworthy observation is the correlation between object density and the complexity of association tasks (Figures 11, 12), implying that higher object densities escalate the intricacies of tracking.

Number of Tracks per Video Figure 10 illustrates the average number of tracks per video. This measure is vital for understanding the range of tracking difficulties, demonstrating that complexity is influenced not only by the quantity of objects but also by the nature of multiple, distinct tracks. High track counts typically indicate more frequent interactions and overlapping paths, posing additional challenges and necessitating sophisticated algorithms for effective differentiation.

The Age and Gender Distributions Figure 8 shows the age and gender distribution of actors in our dataset. By encompassing a wide range of ages and genders, the dataset is relevant to diverse real-world settings. This variety does more than represent different demographics; it introduces added complexity to tracking tasks. Different movement and interaction patterns among various age groups and genders present additional challenges, particularly in dynamic or crowded settings.

5.3. Full Illustrations

We provide full illustrations of the cameras installed in the campus and factory environments in Fig. 13 and Fig. 14, respectively. These cameras were installed in various locations, such as indoors, outdoors, and across different floors, replicating a dense real-world surveillance camera system.

6. Dataset License

The licenses of the datasets used in the experiments are denoted as follows:

- COCO2017 [13]: CC BY 4.0
- Market-1501 [26]: https://zheng-lab.cecs.anu.edu.au/Project/project_reid.html
- MSMT17 [21]: https://www.pkuvmc.com/publications/msmt17.html
- MOT17 [16]: CC BY-NC-SA 3.0
- MOTSynth [6]: CC BY-NC-SA 3.0

7. Video Demo

Our demo highlights the MTMMC dataset's distinct characteristics, including multi-modal data, extended and varied tracks, and complex scenarios. It features diverse situations recorded by RGB-Thermal cameras in campus and factory settings, illustrating the dataset's proficiency in tracking multiple targets across different camera perspectives, managing occlusions, and navigating challenging lighting conditions.

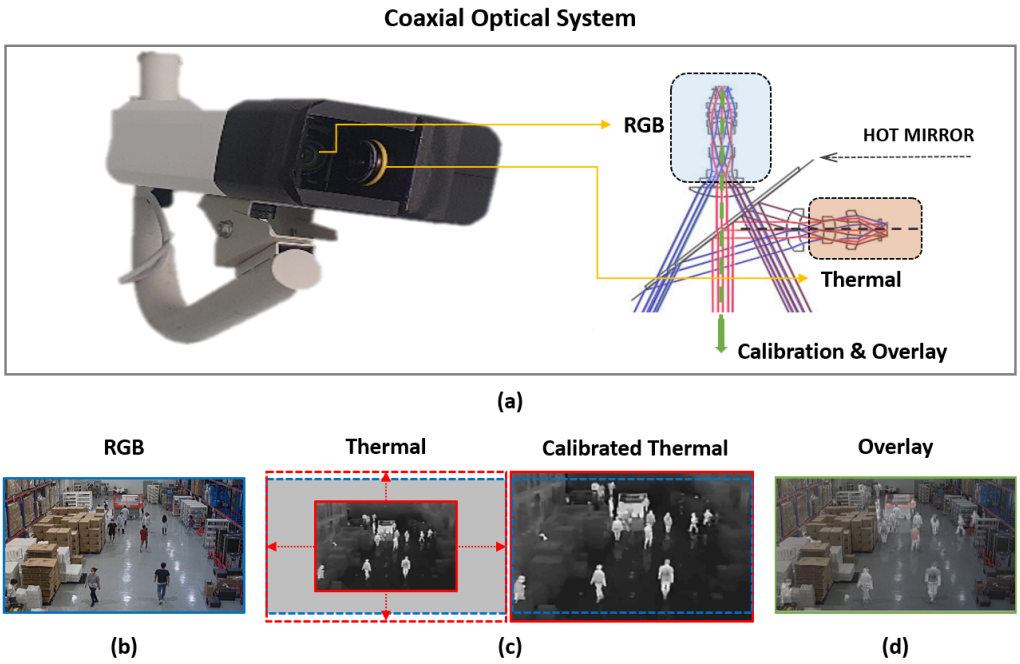


Figure 7. (a) Multi-modal camera with Coaxial Optical System. (b) RGB image. (c) Calibrating Thermal image with RGB image. (d) Overlay image.

	Factory	Campus	Total
trian (#)	7	7	14
val (#)	3	2	5
test (#)	2	4	6
Total (#)	12	13	25

(a) Environment

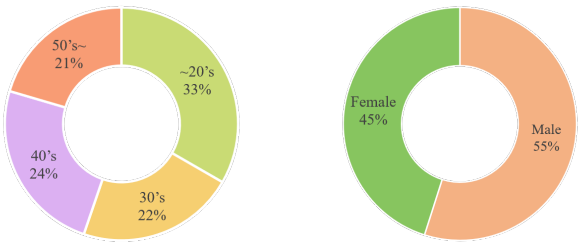
	Sunny	Cloudy	Total
trian (#)	13	1	14
val (#)	3	2	5
test (#)	5	1	6
Total (#)	21	4	25

(b) Weather

	Summer	Fall	Total
trian (#)	7	7	14
val (#)	3	2	5
test (#)	2	4	6
Total (#)	12	13	25

(c) Season

Table 6. **Details on the dataset split.**



(a) Age (b) Gender

Figure 8. **The age and gender distributions of actors.**

8. Ethical Considerations

In creating the MTMMC dataset, our foremost commitment is to the protection of personal privacy and ethical integrity in data usage. We meticulously selected school campuses and factory locations where we had comprehensive authorization, ensuring total control over the data collection process. This approach was governed by a stringent and transparent protocol. Each participant provided informed consent through release agreements, and we rigorously de-identified all non-actor data to further safeguard privacy. This meticulous process received the endorsement of the National Information Society Agency. The Institutional Review Board (IRB) is presently conducting an in-depth review of the dataset, underscoring our dedication to ethical compliance. Furthermore, we are committed to adhering to all relevant privacy laws and regulations, ensuring the dataset aligns with the highest standards of data protection.

Our intention in releasing the MTMMC dataset is to foster advancements in multi-target multi-camera tracking research. We aim to facilitate the generation of open-source, transparent academic works, enhancing knowledge and understanding within the scholarly community. The dataset is strictly designated for non-commercial, public, and academic research, with specific use cases detailed in the accompanying agreement. We vigilantly prohibit any unauthorized use that

diverges from these outlined purposes, particularly to avoid potential civil rights violations. In such instances, we will take decisive action in accordance with applicable laws.

To mitigate any inadvertent misuse by third-party groups, we have instituted robust measures:

- Access to the MTMMC dataset is granted exclusively upon formal request. Prospective users must sign a comprehensive usage agreement, outlining their responsibilities and the ethical boundaries of data utilization.
- We maintain a proactive stance in monitoring dataset usage. The author's affiliated institution(s) reserve the right to report any suspicious activities or individuals to law enforcement officials or regulatory bodies, particularly in cases of legal or regulatory transgressions.
- In collaboration with law enforcement agencies, we will actively participate in investigations and legal actions against any illicit activities involving the dataset.
- Regular audits and reviews of dataset usage will be conducted to ensure continuous adherence to ethical standards and privacy regulations.
- We have established training programs for all dataset users, emphasizing the importance of ethical data handling and awareness of privacy implications.
- A transparent feedback mechanism is in place, allowing users and observers to voice ethical concerns or report misuse. This facilitates a responsive and accountable approach to data governance.
- Our ethical practices are not static; they are subject to ongoing evaluation and refinement, reflecting the evolving landscape of data privacy and ethical norms.
- We engage with external ethical boards and committees, seeking their guidance and oversight in maintaining the ethical integrity of our dataset.

Our comprehensive approach to ethical data management reflects our unwavering commitment to upholding the highest standards of privacy and integrity in academic research. Through these measures, we strive to ensure the MTMMC dataset serves as a valuable and responsible resource for the research community.

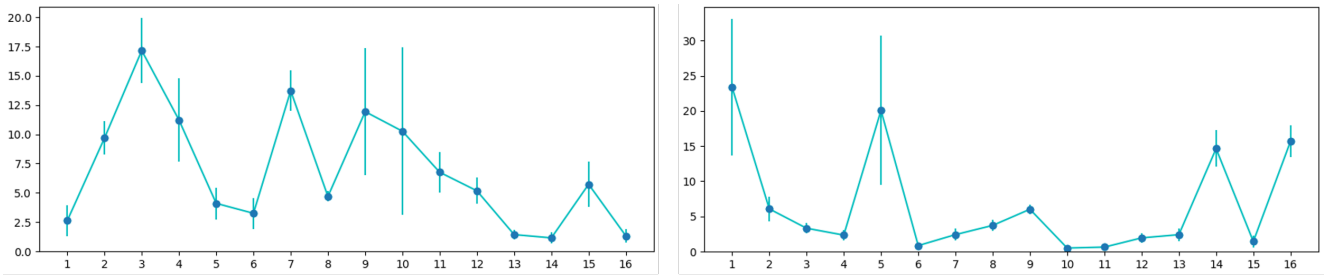


Figure 9. **Number of Objects per Frame.** (a) campus and (b) factory.

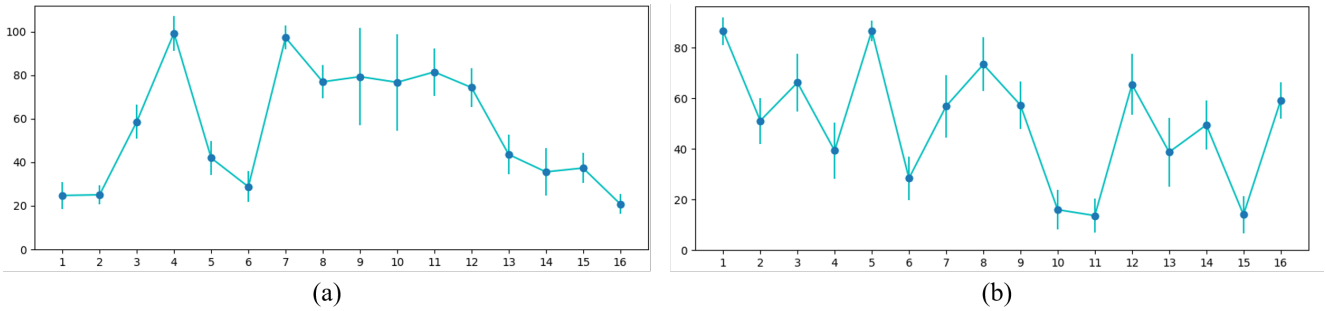


Figure 10. **Number of Tracks per Video.** (a) campus and (b) factory.

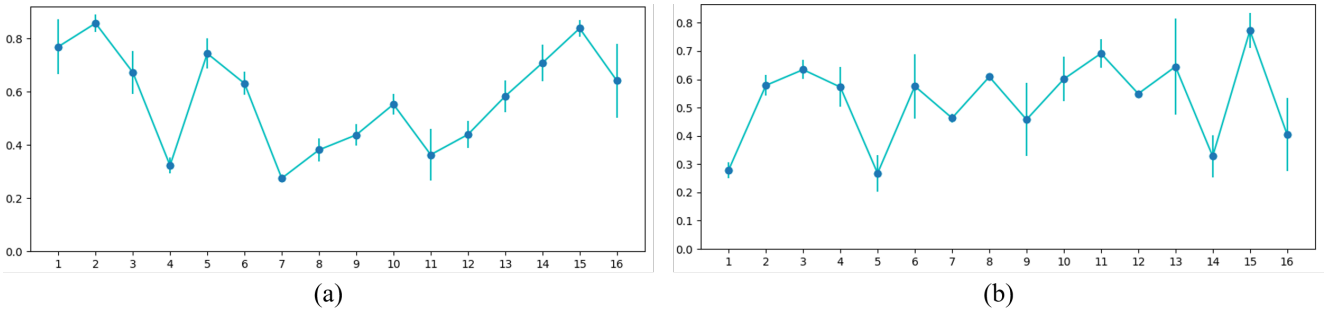


Figure 11. **Analysis on Association Performance.** For the analysis, we adopt AssA [14] with the localisation threshold $\alpha = 50$.

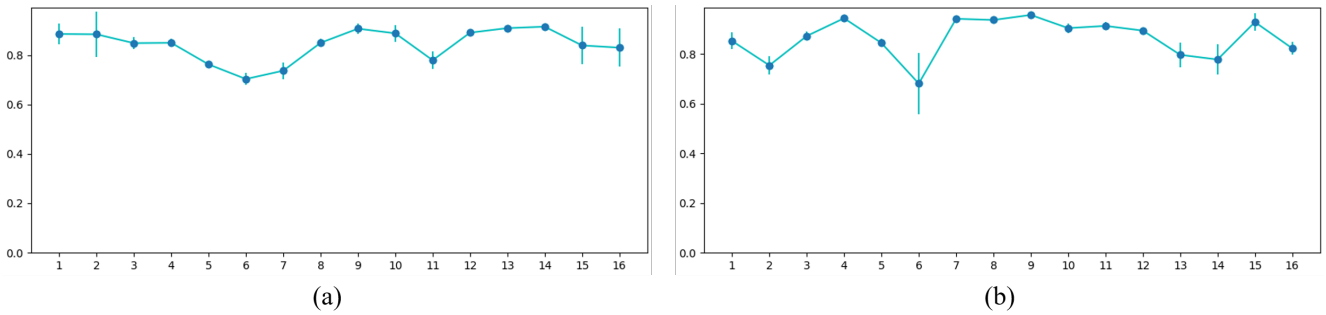
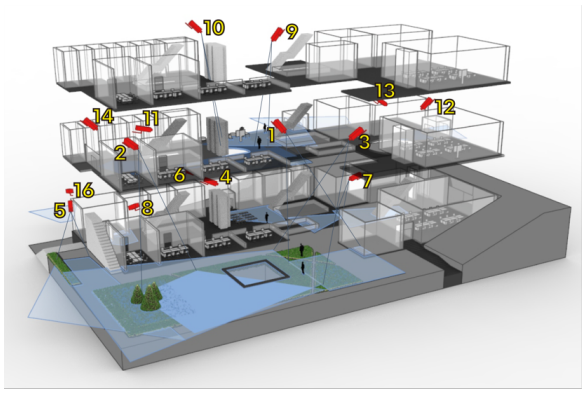
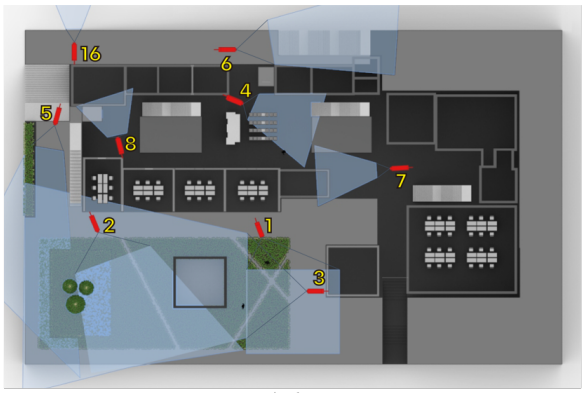


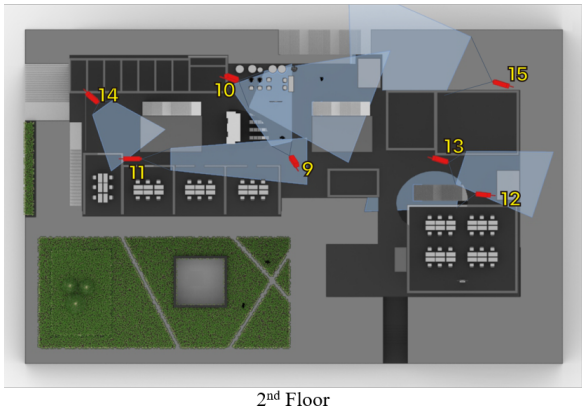
Figure 12. **Analysis on Detection Performance.** For the analysis, we adopt DetA [14] with the localisation threshold $\alpha = 50$.



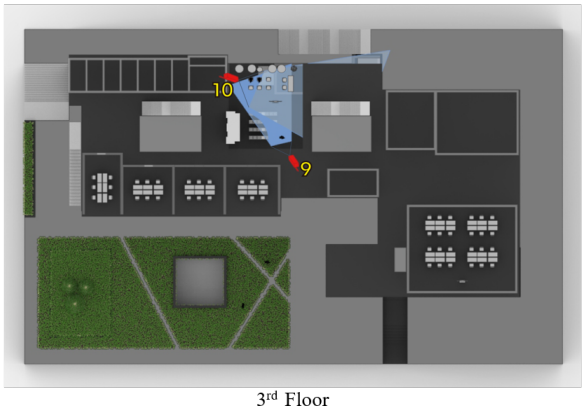
3D-Layout



1st Floor



2nd Floor



3rd Floor

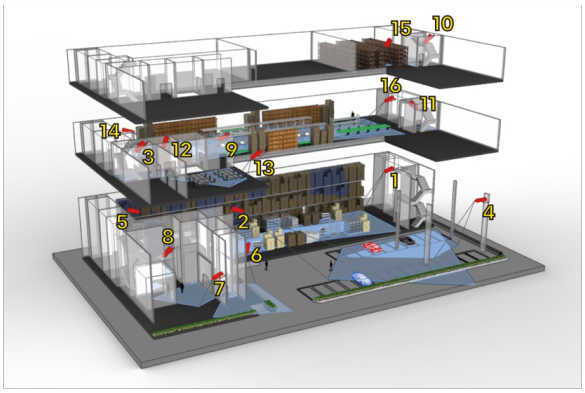
Cam 1 Cam 2 Cam 3 Cam 4

Cam 5 Cam 6 Cam 7 Cam 8

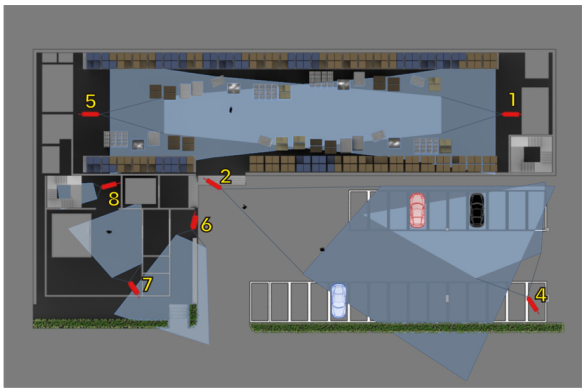
Cam 9 Cam 10 Cam 11 Cam 12

Cam 13 Cam 14 Cam 15 Cam 16

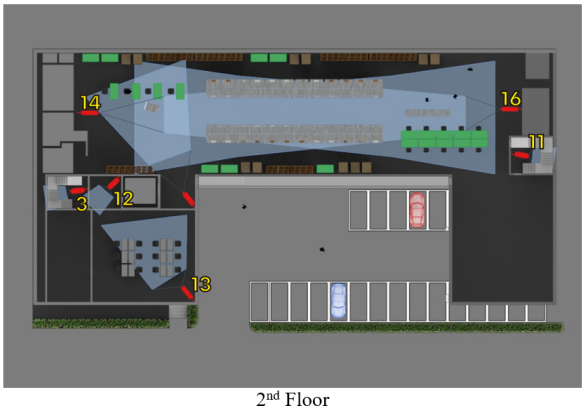
Figure 13. (top) Camera layout on campus environment. (bottom) Examples of multi-spectral images on campus.



3D-Layout



1st Floor

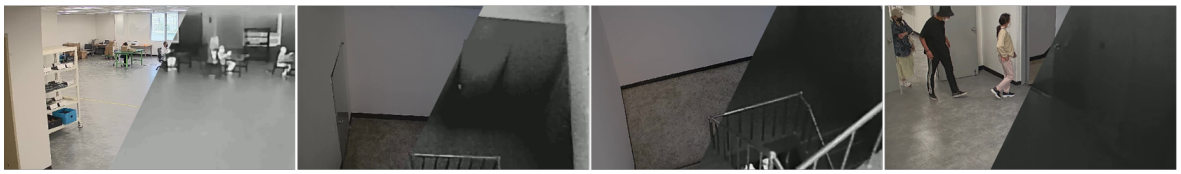


2nd Floor

3rd Floor

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

Figure 14. (top) Camera layout on factory environment. (bottom) Examples of multi-spectral images on factory.

Multi-Target Multi-Modal Camera Tracking Benchmark
Terms and Conditions

First Name	Last Name	
Email Address		
Country		
Address Line 1		
Address Line 2 (Optional)		
City	State	Zip Code
Organization/Affiliation		

Intended Use

Terms and Conditions

Multi-Target Multi-Modal Camera Tracking Benchmark Use Agreement

To access the Multi-Target Multi-Modal Camera Tracking Dataset and any associated materials, text or image files, associated media and "online" or electronic documentation (together, the "**Dataset**"), you (as defined below) must first agree to this Multi-Target Multi-Modal Camera Tracking Dataset Use Agreement ("**Agreement**") . You may not use the Dataset if you do not accept this Agreement. By checking the "I accept the terms and conditions" box below, accessing the Dataset or both, you hereby agree to the terms of the Agreement. If you are agreeing to be bound by the Agreement on behalf of your employer or other entity, you represent and warrant to us that you have full legal authority to bind your employer or such entity to this Agreement. If you do not have the requisite authority, you may not accept the Agreement or access the Dataset on behalf of your employer or other entity.

This Agreement is effective upon the earlier of the date that you first access the Dataset or accept this Agreement ("**Effective Date**"), and is entered into by and between us (or based on where you live, one of its affiliates), and you, or your employer or other entity (if you are entering into this agreement on behalf of your employer or other entity) ("**Participant**" or "**you**").

1. Subject to Participant's compliance with the terms and conditions of this Agreement, We permit Participant to: (a) use the **Dataset**, including the data and the annotations, for non-commercial, research, or academic purposes only to research, develop and improve software, algorithms, machine learning models, techniques and technologies designed to train and evaluate AI and machine-learning models for multi-target multiple camera tracking tasks (the "**Purpose**"); (b) for analyzing and testing purposes; and (c) publish (or present papers or articles) on your results from using the Dataset, provided that no material portion of the Dataset is included in any such publication or presentation; provided, however, you are permitted to distribute and reproduce up to ten (10) minutes of video from the Dataset per Participant research or academic publication related to the Purpose. You shall implement and maintain appropriate technical and organizational data protection and security measures to ensure security of the Dataset, including without limitation the measures necessary to protect against unauthorized or unlawful access, acquisition or use of the Dataset and against accidental loss, destruction or damage of or to the Dataset.
2. Subject to Participant's compliance with the terms and conditions of this Agreement, Participant retains its intellectual property rights in and to all algorithms, software, machine learning models, techniques and technologies developed or otherwise derived by Participant from the use of the Dataset. Such algorithms, software, machine learning models, techniques and technologies can only be used for academic purposes.
3. As between us and Participant, we retain all intellectual property rights and all other rights, title, and interest in and to the Dataset. You acquire no interest in the Dataset you receive under the terms of this Agreement. All rights not expressly granted under this Agreement by us are reserved.
4. At any time, we may require Participant to delete all copies of the Dataset (in whole or in part) in Participant's possession and control. Participant will promptly comply with any and all such requests. Upon our request, Participant shall provide us with written confirmation of Participant's compliance with such requirement.
5. If we reasonably believe that you are or are likely to be in violation of the terms of this Agreement, our designee may audit your use, storage and distribution of the Dataset, including, without limitation, any and all records and files associated with the Dataset and this Agreement. You hereby agree to cooperate with such audit.
6. Participant will not:
    A. distribute, copy, disclose, assign, sublicense, embed, host or otherwise transfer the Dataset to any third party, except as described in Section 1(b) above;
    B. remove or alter any copyright, trademark or other proprietary notices appearing on or in copies of the Dataset;
    C. use the provided trademarks in a way that suggests publications or presentations come from or are endorsed by us;

D. use the Dataset to measure, detect, predict, or otherwise label the race, ethnicity, age, or gender of individuals;

E. use the Dataset to extract or process biometric identifiers or biometric information;

F. use the Dataset in a pornographic, defamatory, malicious, deceptive or unlawful manner, or in violation of any applicable regulations or laws (including applicable data protection and privacy law);

G. incorporate the Dataset into any other program, dataset, or product;

H. use the Dataset to distribute images or videos (except as expressly set forth in Section 1(b) above); or

I. use the Dataset for any purpose other than the Purpose specified in this Agreement.

7. If you use the Dataset (or any portion thereof) in a manner that features models or property in connection with a subject that would be unflattering or unduly controversial to a reasonable person, you must indicate: (1) that the content is being used for illustrative purposes only, and (2) any person depicted in the content is a model. For example, you could say: "Stock photo. Posed by model."

8. If you give feedback about the Dataset to us, you give us, without charge, the right to use, share and commercialize your feedback in any way and for any purpose. You also give to third parties, without charge, any patent rights needed for their products, technologies and services to use or interface with any specific parts of our dataset or service that includes the feedback. You will not give feedback that is subject to a license that requires us to license its Dataset or documentation to third parties because we include your feedback in them. These rights survive this Agreement.

9. Upon the termination of this Agreement, Participant will immediately stop using the Dataset and destroy all copies of the Dataset and related materials in Participant's possession and control. Additionally, we may, at any time, for any reason or for no reason, terminate this Agreement, effective immediately upon notice to the Participant. Upon termination, the license granted to Participant hereunder will immediately terminate and Participant will immediately stop using the Dataset and destroy all copies of the Dataset and related materials in Participant's possession or control. Except for the licenses and rights granted to Participant, the other provisions of this Agreement will survive any termination.

10. THE DATASET IS PROVIDED "AS IS" WITHOUT ANY EXPRESS OR IMPLIED WARRANTY OF ANY KIND, INCLUDING, WITHOUT LIMITATION, WARRANTIES OF MERCHANTABILITY, TITLE, NON-INFRINGEMENT, OR FITNESS FOR ANY PARTICULAR PURPOSE.

11. IN NO EVENT WILL AUTHOR'S INSTITUTION AND ITS CONTRACTORS BE LIABLE FOR ANY DIRECT, CONSEQUENTIAL, INCIDENTAL, EXEMPLARY, PUNITIVE, SPECIAL, OR INDIRECT DAMAGES (INCLUDING DAMAGES FOR LOSS OF PROFITS, BUSINESS INTERRUPTION, OR LOSS OF INFORMATION) ARISING OUT OF OR RELATING TO THIS AGREEMENT OR ITS SUBJECT MATTER, EVEN IF AUTHOR'S INSTITUTION HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

12. WE AND ITS CONTRACTOR'S TOTAL LIABILITY ARISING FROM OR RELATING TO THIS AGREEMENT AND ITS SUBJECT MATTER WILL NOT EXCEED ONE HUNDRED DOLLARS ($100).
13. Either party may terminate this Agreement if the other is in material breach of this Agreement and such breach remains uncured for thirty (30) days following receipt of written notice of the breach.
14. Participant will comply with all applicable export controls, import controls and trade sanctions applicable to the Dataset. You shall obtain, at your sole cost and expense, any export and import (temporary and permanent) license and other official authorization applicable to the Dataset.
15. You will defend, indemnify and hold Author's Insitution, including its subsidiaries, affiliates and agents (collectively the "**Indemnified Parties**") harmless from all expenses (including all judgments, settlements, attorneys' fees and costs) related to any claim or action arising from or by reason of your failure to comply with the terms of this Agreement. The Indemnified Party will: (1) promptly notify the indemnifying party of any claim or action, (2) permit the indemnifying party (through mutually-agreed counsel) to answer and defend the claim or action, and (3) provide non-confidential information and assistance, at the indemnifying party's expense and request, as needed to answer and defend the claim or action. The indemnifying party may not settle or publicize any claim or action without the Indemnified Party's consent.

☐ I accept the terms and conditions

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 6

[2] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 6

[3] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 3

[4] Liuyuan Deng, Ming Yang, Tianyi Li, Yuesheng He, and Chunxiang Wang. Rfbnet: deep multimodal networks with residual fusion blocks for rgb-d semantic segmentation. *arXiv preprint arXiv:1907.00135*, 2019. 4

[5] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 4

[6] 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. 3, 6

[7] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010. 3

[8] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei. Fastreid: A pytorch toolbox for general instance re-identification. *arXiv preprint arXiv:2006.02631*, 2020. 3

[9] 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. 5

[10] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015. 6

[11] 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. 5

[12] 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. 4

[13] 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. 3, 6

[14] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021. 9

[15] 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. 5, 6

[16] 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. 3, 6

[17] 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. 4, 5

[18] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018. 4

[19] 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. 3

[20] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. Deepsportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, pages 1–8, 2022. 5

[21] 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. 3, 5, 6

[22] 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. 3

[23] 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. 4

[24] 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, 6

[25] 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. 4

[26] 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 con-*

ference on computer vision, pages 1116–1124, 2015. 3, 5, 6

[27] Wujie Zhou, Xinyang Lin, Jingsheng Lei, Lu Yu, and Jenq-Neng Hwang. Mffenet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing. *IEEE Transactions on Multimedia*, 24:2526–2538, 2021. 4