Visio-Temporal Attention for Multi-Camera Multi-Target Association

Yu-Jhe Li Xinshuo Weng Yan Xu Kris Kitani
Carnegie Mellon University
{yujheli, xinshuow, yxu2, kkitani}@cs.cmu.edu

Abstract

We address the task of Re-Identification (Re-ID) in multi-target multi-camera (MTMC) tracking where we track multiple pedestrians using multiple overlapping uncalibrated (unknown pose) cameras. Since the videos are temporally synchronized and spatially overlapping, we can see a person from multiple views and associate their trajectory across cameras. In order to find the correct association between pedestrians visible from multiple views during the same time window, we extract a visual feature from a tracklet (sequence of pedestrian images) that encodes its similarity and dissimilarity to all other candidate tracklets. We propose a inter-tracklet (person to person) attention mechanism that learns a representation for a target tracklet while taking into account other tracklets across multiple views. Furthermore, to encode the gait and motion of a person, we introduce second intra-tracklet (person-specific) attention module with position embeddings. This second module employs a transformer encoder to learn a feature from a sequence of features over one tracklet. Experimental results on WILDTRACK and our new dataset 'ConstructSite' confirm the superiority of our model over state-of-the-art ReID methods (5% and 10% performance gain respectively) in the context of uncalibrated MTMC tracking. While our model is designed for overlapping cameras, we also obtain state-of-the-art results on two other benchmark datasets (MARS and DukeMTMC) with non-overlapping cameras.

1. Introduction

Multi-Target Multi-Camera (MTMC) tracking [22, 37] relies deeply on the ability to associate people between multiple cameras to determine the position of each person over time. Depending on the situations, the cameras may be synchronized, calibrated (known position) or have overlapping views. In this work, we focus on the case where cameras are synchronized with overlapping views, but the calibration information is not available. Our aim is to develop a method that can perform association of pedestrian trajectories across cameras without using any calibration information about the cameras (shown in Figure 1).

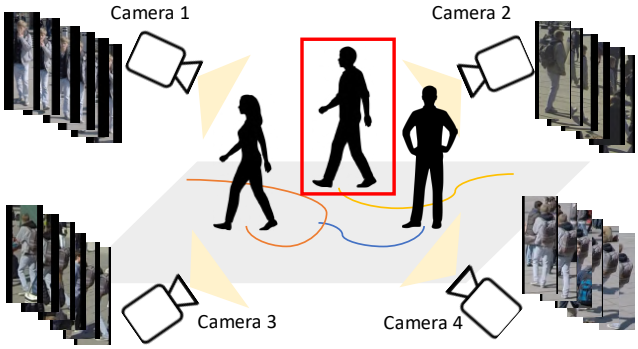

Figure 1: **Multi-target multi-camera tracking with overlapping views.** When the target person is seen from multiple synchronized cameras (views), identifying the person is feasible by finding the similarities and dissimilarities across multiple views. Note that the geometry information such as the position of each camera may not be known.

In order to develop such a method for data association across uncalibrated cameras, we need to extract a discriminative feature for each person over a sequence of image patches (tracklet) and perform feature matching across tracklets in different cameras. This process we have described is a form of the Re-Identification (Re-ID) problem [71]. In the time-synchronized MTMC scenario, the Re-ID problem is simplified since we only need to match pedestrians appearing in multiple cameras during the same time window. Within this time window, we would like to extract visual features that are both representative and discriminative (sufficiently different from other pedestrians in the same view) so that we can match people across camera views.

To learn both a representative and discriminative visual feature for robust person association across views, we propose a novel video-based Re-ID model using Transformers [50]. Since attention models [63] has the ability to learn and embed the similarity and dissimilarity between different synchronized tracklets from overlapping views, it can be used to learn representative and discriminative visual fea-

tures. We use attention models in two ways: (1) we introduce an inter-tracklet attention model to learn the correlation between tracklets across cameras and (2) we introduce an intra-tracklet attention module (before the inter-tracklet attention model) to learn a person-specific motion and appearance feature.

In order to evaluate our Re-ID method for MTMC tracking, we use a construction site dataset (which we call ConstructSite) provided by a construction company. Videos in the dataset are recorded in a construction site with unknown camera positions. Recorded with four synchronized cameras, this dataset has 88 videos (3-minute long) where each synchronized camera has 22 videos. As mentioned above, our Re-ID method is designed intentionally for overlapping, time synchronized, uncalibrated cameras. We also perform experiments on two other public benchmark datasets (MARS and DukeMTMC) with non-overlapping cameras. The contributions of this paper are highlighted below:

1. We introduce a transformer-based inter-tracklet attention module that computes a discriminative feature representation by taking into account all other time synchronized tracklets across all camera views.
2. In order to learn a person-specific motion and appearance feature, we introduce an transformer-based intra-tracklet attention module to learn a compact representation for each tracklet.
3. We show superior Re-ID performance in the time synchronized uncalibrated setting. Furthermore, we apply our method to the case of non-overlapping cameras. We show how our method is able to generalize to harder scenarios while also advancing the state of the art.

2. Related Works

Re-Identification (Re-ID). Re-ID can be categorized into image-based and video-based methods. Image-based person Re-ID [3, 6, 7, 19, 25, 26, 27, 29, 39, 41, 43, 46, 54, 55] typically focuses on matching images with viewpoint and pose variations, or those with background clutter or occlusion. Most of video-based methods use optical flow [5, 9, 32, 64], recurrent neural networks (RNNs), temporal pooling [69], or spatial-temporal attention to model the temporal information. On the other hand, several attention-based methods [25, 41, 43] are further proposed to focus on learning the discriminative image features. Compared with temporal pooling [69] which assigns the same weights to all frames, prior attention-based methods [13, 23, 31, 64, 73] learn the weight of different frames or parts from a static perspective, i.e. considering the spatial attention and temporal attention separately. Yet, in comparison to prior work regarding attention models (mostly with single-head self-attention), our developed model using transformers is able to learn more discriminative features (person to person and person-specific) with a series of multi-head self-attention modules. This is well adapted to the MTMC tracking with overlapping camera views.

Multi-target multi-camera (MTMC) tracking. In terms of MTMC tracking, one has to address two distinct but closely related research problems: 1) Detection and tracking of targets within a single camera, known as single camera tracking(SCT); 2) Re-ID of targets across multiple cameras. That is, MTMC tracking can be regarded as the combination of SCT within cameras and Re-ID with spatial-temporal information to connect target trajectories across cameras. While previous Re-ID work achieves promising performance, adapting Re-ID into the MTMC tracking pipeline is a challenging task. With the recent development in Re-ID, a number of MTMC tracking methods [22, 30, 37, 66, 68] adopting Re-ID technology have been proposed. In [37], Ristani *et al*. learn the feature for both MTMC tracking and Re-ID with a convolutional neural network. In [68], Zhang *et al*. obtain promising results with simple hierarchical clustering and Re-ID feature. In [22], Li *et al*. utilize occlusion and orientation status in the Re-ID model which leads to an improved MTMC tracking performance. However, few of the recent Re-ID works apply MTMC tracking with overlapping cameras. As a complement, our work demonstrates the use of Re-ID model for MTMC tracking in both overlapping and non-overlapping scenarios.

Single-camera multi-object tracking (SCT). With the advancements of object detection, tracking-by-detection framework is widely used in single-camera multi-object tracking, where the detection module is followed by data association across frames. To solve the data association problem, prior work can be categorized into offline and online methods. Offline methods [1, 10, 38, 40, 47, 48, 49, 51, 56] attempt to adopt global optimization with the access of data of the entire sequence. On the other hand, online methods [2, 8, 14, 15, 42, 45, 53, 57, 58, 59, 60, 62, 67] solve data association given only data up to the current frame. As the focus of this paper is to improve the Re-ID model for MTMC tracking across cameras, we use a baseline SCT approach – DeepSort [59] to formulate tracklets in each camera for simplicity. Although, the single-camera tracking approach used in our pipeline can be replaced with other approaches.

3. Dataset

3.1. Previous datasets

Current person Re-ID datasets have significantly pushed forward the research on person Re-ID. As shown in Table 1, MSMT17 [55], DukeMTMC-reID [35, 72], CUHK03 [24], and Market1501 [70] involve large numbers of cameras and identities. The extended datasets from Market1501 and

Table 1: Publicly available benchmarks for person image-signature-based Re-ID, video-based MTMC tracking, and MTMC Overlap datasets with overlapping cameras. We only list commonly used datasets.

	Dataset	# cameras	Overlapping	Video	Geometry	#bboxes	#IDs	Target
Re-ID	DukeMTMC-reID [35, 72]	8	✗	✗	✗	36,411	1,812	pedestrian
	Market1501 [70]	6	✗	✗	✗	32,668	1,501	pedestrian
	MSMT17 [55]	15	✗	✗	✗	126,441	4,101	pedestrian
	CUHK03 [24]	5	✗	✗	✗	14,097	1,467	pedestrian
MTMC-nonoverlap	MARS [69]	6	✗	✓	✗	1,191,003	1,261	pedestrian
	DukeMTMC [35, 61]	8	✗	✓	✗	126,441	1,812	pedestrian
MTMC-overlap	Laboratory [12]	4	✓	✓	✓	476	6	pedestrian
	Terrace [12]	4	✓	✓	✓	1,023	9	pedestrian
	Passageway [12]	4	✓	✓	✓	226	13	pedestrian
	Campus [65]	4	✓	✓	✓	240	25	pedestrian
	WILDTRACK [4]	7	✓	✓	✓	66,626	313	pedestrian
	ConstructSite (ours)	4	✓	✓	✗	4,806,564	440	worker

DukeMTMCreID are also available for video-based Re-ID and MTMC, which are MARS [69] and DukeMTMC [35, 61], respectively. Though the trajectory information is available in MARS, the original videos and camera geometry are unknown to the public. In contrast, DukeMTMC provides camera network topology so that the relative location between cameras among cameras can be established. However, the cameras in DukeMTMC and MARS are non-overlapping. Again, as "overlap" datasets we refer to the ones whose camera's fields of view strictly overlap. The three sequences shot at the EPFL campus [12]: Laboratory, Terrace, and Passageway, as well as Campus [65] are overlapping multi-camera datasets. These four datasets have a small number of total identities and are relatively sparsely crowded. As we can see from Table 1, Laboratory, Terrace, Passageway, and Campus are of small size. WILDTRACK [4] improves upon other overlapping datasets as it has a larger number of annotated identities that allow for developing deep learning-based MTMC approaches.

3.2. ConstructSite

In order to evaluate our model in a different scenario (not campus with pedestrians walking as in prior datasets), we develop a new MTMC dataset called *ConstructSite*. This dataset contains 88 videos, each of which with 3 minutes long. *ConstructSite* is captured by 4 synchronized cameras (each camera has 22 videos). These videos are recorded in a construction site, where workers are *wearing work-wears* instead of casual-wears and performing *a variety of actions* in addition to walking such as squating, kneeling, carrying. Some examples from the dataset are presented in the Figure 2. We detail the information on hardware and annotation as below.

Hardware. The dataset is recorded using 4 statically positioned HD cameras. In particular, we use four *GoPro Hero 7 cameras* to record the videos and downsample the videos into resolution 1352×760 pixels with the frame rate of 30 (FPS). The synchronization accuracy between the four cameras is about 100 ms.

Figure 2: **Examples of workers in the *ConstructSite*.** The action type of each worker spans walking, standing, squating, kneeling, and etc.

Annotation. As stated earlier, there are 88 videos totally in the dataset while each camera has 22 videos. That is, there are 22 synchronized videosets. There are around 15 to 20 unique identities (IDs) of workers for each videoset and around 50,000 bounding-boxes for each video. We would like to note that, all of the bounding-boxes in this dataset are manually labelled. This results in 4,806,564 bounding-boxes, each of which is with an associated ID.

4. Approach

To achieve MTMC tracking in the overlapping scenario, we propose our video-based Re-ID model to associate the tracklets across cameras during the same time window. The input tracklets can be ground-truth tracklets in each cam-

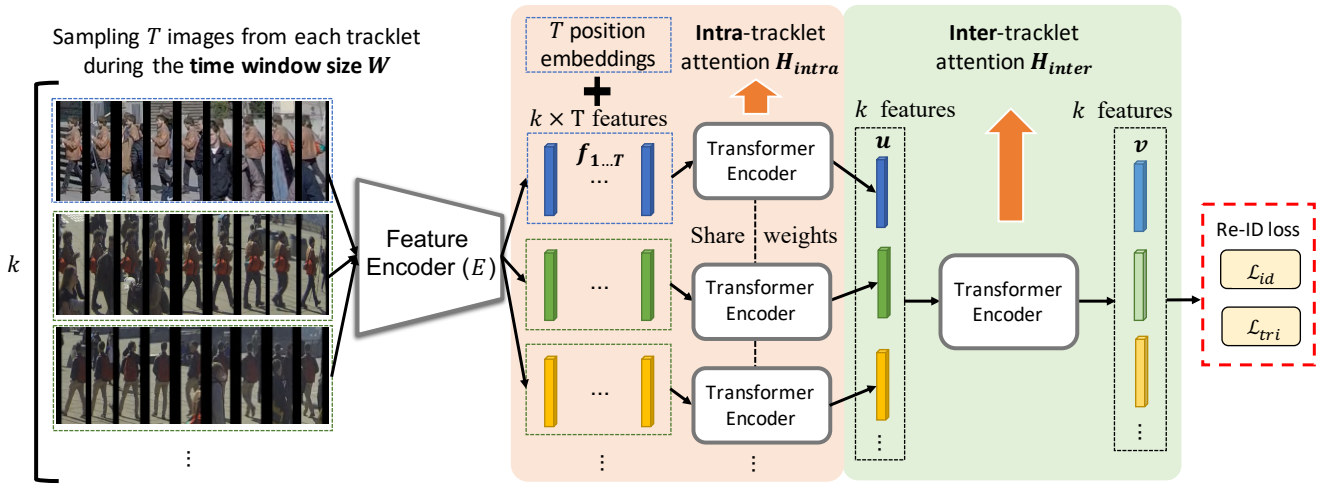

Figure 3: **Our proposed video-based Re-ID model for MTMC tracking**. Our model is composed of three modules: the feature encoder E, intra-tracklet attention module H_{intra}, and inter-tracklet attention module H_{inter}. To extract the visual features from the k tracklet sets each with T sampled images, we apply the feature encoder E to obtain the $k \times T$ features. With the use of intra-tracklet attention module H_{intra}, we derive the newly attended $k \times T$ features within tracklets. Finally, with the use of inter-tracklet attention module H_{inter}, we are able to derive the k representative features across these k tracklets during the same time window.

era but without cross-camera association or output tracklets from a real-world single-camera tracking method. Given a cropped image sequence (tracklet) of a pedestrian, our goal is to learn a model to extract a representative and discriminative feature representation that enables video-based person Re-ID across cameras. Specifically, we have a tracklet with the sampled frame set $X = \{x_i\}_{i=1}^{T}$ along with the associated label y, where $x_i \in \mathbb{R}^{H \times W \times 3}$ and $y \in \mathbb{N}$ represent the identity for this tracklet. There are several ways to sample these T frames from a tracklet in the window size W in order to handle the long-range temporal structure. To balance speed and accuracy, we adopt the restricted random sampling strategy [23, 52].

As shown in Figure 3, our model is composed of three modules: (1) the feature encoder E; (2) the intra-tracklet attention module H_{intra}, and (3) the inter-tracklet attention module H_{inter}. First, to extract the visual features from the tracklet set X, we apply the feature encoder E and obtain a feature set $F = \{f_i\}_{i=1}^{T}$ for one tracklet, where $f_i \in \mathbb{R}^d$ (d denotes the dimension of the visual feature). Second, the intra-tracklet attention module H_{intra} takes F with T sequential features and T learnable position embeddings as input and produces a visual feature u representing the tracklet, where $u \in \mathbb{R}^d$ (d denotes the dimension of the feature). Third, the inter-tracklet attention module H_{inter} takes $U = \{u_i\}_{j=1}^{k}$ from all k tracklets and produces the updated features $V = \{v_i\}_{j=1}^{k}$ for these tracklets. The backbones of both H_{intra} and H_{inter} employ transformer encoders. With the joint use of intra-tracklet and inter-tracklet attention modules, our model produces a representative feature for each tracklet during the window size W. We will detail each attention module in the following sections.

To perform video-based Re-ID across cameras in the testing phase, our framework encodes each tracklet into a representation, which is later applied for matching the nearest ones via nearest neighbor search for Re-ID. We additionally use Hungarian algorithm [20] for MTMC tracking during inference scenario.

4.1. Preliminary

The transformer encoder (shown in Figure 4) used in both of our attention modules is inspired by the Transformer [50], which features a series of encoders and decoders of an identical structure. Every encoder has a multi-head self-attention layer (MHSA) and a feedforward network layer.

Standard self-attention. For the sake of completeness, we briefly review the self-attention module [63]. A typical self-attention layer transforms the input features into three inputs: query Q, key K, and value V by matrix multiplication with transforming matrix. The softmax layer will take the result of the multiplication of Q and K, and produce the attention weights. The target attention result is then produced from the result of the final matrix multiplication of softmax and the V.

Multi-head self-attention. To observe both temporal and concept information from the input features $Z = \{z_m\}_{m=1}^{M}$, we advance the idea of multi-head self-attention. As depicted in Fig 4, we have the entire attention module H comprising of N self-attention modules (i.e., the head number equals N), and each of them is developed to derive the attention features in N subspaces. We first transform the in-

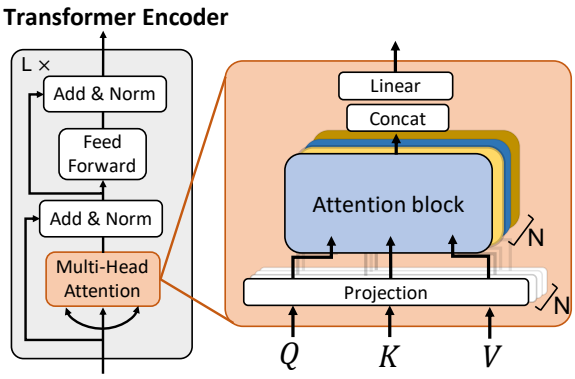

Figure 4: Illustration of multi-head attention in the transformer encoder. With N different single-head attention blocks (each with a projection matrix layer), self-attention can be performed in different subspaces (dimension d for each) for capturing diverse visual concepts. We concatenate the outputs $O_{1:N}$ from all attention blocks and obtain the joint attention result at the output of the final linear transform layer.

put W into N subspaces using the N projection layers M_n ($\mathbb{R}^{d_n} \leftarrow \mathbb{R}^d$) where n denotes the projection layer number ($n = 1 \sim N$) and d_n denotes the subspace dimension. To produce the finalized results from all of the N subspaces, the introduced a linear projection layer M^R to derive the final attended features $R = \{r_m\}_{m=1}^{M}$, where $r_k \in \mathbb{R}^d$ (same dimension as the original input features $Z = \{z_m\}_{m=1}^{M}$). The above procedure can be formulated as:

$$R = M^R \cdot \text{concat}(O_{1:N}),$$

where concat means we concatenate the outputs $O_{1:N}$ from all of the N self-attention blocks.

4.2. Intra-tracklet attention

To capture feature of the entire motion sequence of a person and appearance temporally, we introduce the intra-tracklet attention module H_{intra} using transformer encoders to learn a representative feature of each tracklet. Following [11, 34], we employ the standard learnable 1D position embeddings denoted: $P = \{p_i\}_{i=1}^{T}$ for each of the input T visual features. The intra-tracklet attention module H_{intra} takes $F = \{f_i\}_{i=1}^{T}$ with T sequential features and T learnable position embeddings as input, which can be denoted as:

$$F' = F + P. \tag{1}$$

It then produces a visual feature u representing the tracklet, where $u \in \mathbb{R}^d$ (d denotes the dimension of the visual feature):

$$u = H_{\text{intra}}(F'). \tag{2}$$

Since we have k tracklets during the given time window size W, we will have k feature set as: $U = \{u_j\}_{j=1}^{k}$ from the feature set: $\{F_j\}_{j=1}^{k}$.

4.3. Inter-tracklet attention

Furthermore, to learn and embed the similarity and dissimilarity between all these synchronized tracklets from overlapping views, we further apply the inter-tracklet attention module H_{inter} and derive the finalized representation for each tracklet. Specifically, the inter-tracklet attention module H_{inter} takes $U = \{u_i\}_{j=1}^{k}$ from all k tracklets and produces updated features $V = \{v_i\}_{j=1}^{k}$. That is,

$$V = H_{\text{inter}}(U). \tag{3}$$

Note that, the introduced inter-tracklet attention module H_{inter} aims to attend at other tracklets from multiple views during the same time window. That is, the updated visual features will depend on other tracklets locally, which, as a result, helps the data association in the tasks of Re-ID and MTMC tracking. The produced feature v for each tracklet will finally be used for matching across multiple views.

4.4. Full objective

To better utilize the label information to update our entire network, we first employ classification loss on the output feature vector w, by computing the negative log-likelihood between the predicted label $\tilde{y} \in \mathbb{R}^K$ and the ground truth one-hot vector $\hat{y} \in \mathbb{N}^K$. The identity loss $\mathcal{L}_{id}$ can be represented as

$$\mathcal{L}_{id} = -\mathbb{E}_{(x,y)\sim(X,Y)} \sum_{k=1}^{K} \hat{y}_k \log(\tilde{y}_k), \tag{4}$$

where K is the number of identities (classes). To further enhance the discriminative property, we impose a triplet loss $\mathcal{L}_{tri}$, which aims to maximize the inter-class discrepancy while minimizing intra-class distinctness. Specifically, for each input image x, we sample a positive image x_{pos} with the same identity label and a negative image x_{neg} with different identity labels to form a triplet. The distances between x and x_{pos}/x_{neg} can be computed as:

$$d_{\text{pos}} = \|v_x - v_{x_{\text{pos}}}\|_2, \tag{5}$$

$$d_{\text{neg}} = \|v_x - v_{x_{\text{neg}}}\|_2, \tag{6}$$

where v_x, $v_{x_{\text{pos}}}$, and $v_{x_{\text{neg}}}$ represent the feature vectors of images x, x_{pos}, and x_{neg}, respectively. We then have the triplet loss $\mathcal{L}_{tri}$ defined as

$$\mathcal{L}_{tri} = \mathbb{E}_{(x,y)\sim(X,Y)} \max(0, m + d_{\text{pos}} - d_{\text{neg}}), \tag{7}$$

where $m > 0$ is the margin used to define the difference between the distance of positive image pair d_{pos} and the distance of negative image pair d_{neg}. Hence, the total loss $\mathcal{L}$ for training our proposed network is summarized as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{tri}}. \tag{8}$$

Table 2: **Comparison of video-based Re-ID on the ConstructSite, WILDTRACK, and DukeMTMC dataset.** The number in bold represents the best result. * indicates the code is not released or available.

Method	Source	ConstructSite			WILDTRACK			Mars			DukeMTMC		
		Synchronized & overlapping						Non-synchronized & non-overlapping					
		Rank1	Rank5	mAP	Rank1	Rank5	mAP	Rank1	Rank5	mAP	Rank1	Rank5	mAP
ResNet-50 [18]	CVPR16	69.7	94.4	73.1	70.4	88.9	57.5	84.3	93.8	79.1	94.5	98.3	92.7
ETAP-Net [61]	CVPR18	72.3	93.4	71.3	71.2	88.7	58.4	80.8	92.1	67.4	83.6	94.6	78.3
STA [13]*	AAAI19	-	-	-	-	-	-	86.3	95.7	80.8	96.2	99.3	94.9
GLTR [21]	ICCV19	78.2	94.6	75.1	75.8	89.2	59.7	87.0	95.8	78.5	96.2	99.3	93.7
TKP [17]	ICCV19	77.2	94.0	73.8	77.6	91.3	59.6	84.0	93.7	73.3	94.0	-	91.7
COSAM [44]	ICCV19	76.3	94.4	73.1	77.5	91.2	59.3	84.9	95.5	79.9	95.4	99.3	94.1
NVAN [28]	BMVC19	85.0	95.4	78.0	80.4	92.6	66.3	90.0	-	82.8	96.3	-	94.9
VKD [33]	ECCV20	85.6	96.0	80.1	79.9	92.5	66.1	89.4	96.8	83.1	95.2	98.6	93.5
AP3D [16]	ECCV20	85.4	95.8	80.5	80.3	92.1	67.0	90.1	-	**85.5**	96.3	-	**95.6**
Ours ($L=1$)	default	94.2	99.1	90.8	85.1	96.5	71.6	90.2	96.5	83.2	95.7	99.1	94.9
Ours ($L=3$)		**94.7**	**99.3**	91.0	**85.5**	96.8	**72.0**	**91.4**	97.0	83.8	96.4	**99.4**	95.2
Ours ($L=5$)		94.5	99.2	**91.1**	85.4	**96.9**	71.7	90.5	**97.2**	84.2	**96.5**	**99.4**	95.3

The entire framework which includes feature encoder E, intra-tracklet attention H_{intra}, and inter-tracklet attention H_{inter} is trained and updated end-to-end using this loss.

5. Experiments

5.1. Datasets

To evaluate our Re-ID method, we conduct experiments on two datasets with overlapping cameras: our ConstructSite and WILDTRACK [4], and two benchmark datasets with non-overlapping cameras: MARS [69] and DukeMTMC-VideoReID [35, 61].

ConstructSite. Details of ConstructSite can be found in Section 3.2. In addition, we split the 22 videosets (88 videos) into two halves for training/testing, and each split has 11 videosets (44 videos). For training and testing purposes, we prepare the ground-truth associated tracklets for each camera.

WILDTRACK [4]. The Wildtrack dataset includes 400 synchronized frames from 7 cameras. These 7 cameras capture images of pedestrians, and the bounding boxes are annotated 2 frames per second (fps). The dataset has 313 identities of pedestrians entirely, and we split the first 250 for training and the remaining 63 for testing after we crop the person images accordingly.

MARS [69]. MARS is a large-scale video-based person re-identification benchmark dataset with 17,503 sequences of 1,261 identities and 3,248 distractor sequences. The training set contains 625 identities, and the testing set contains 636 identities

DukeMTMC-VideoReID [61]. The DukeMTMC-VideoReID dataset is another large-scale benchmark dataset with 4,832 tracklets of 1,812 identities for video-based person Re-ID. It is derived from the DukeMTMC dataset [35]. The dataset is divided into 408, 702 and 702 identities for distraction, training, and testing, respectively.

5.2. Implementation details

We resize each cropped person image to 256×128 in MARS, DukeMTMC, and WILDTRACK, while to 224×224 (square) for ConstructSite. This is due to the fact that images of person in ConstructSite have several other actions such as squating down or cross-legged sitting. The sampling number T is set as 8 for each tracklet following [23, 52]. The window size for MTMC tracking is set as W=30. We use ResNet-50 pre-trained on ImageNet as our backbone of feature encoder E. The other two attention modules are composed of L-layer of transformer encoders while L is selected as 3. We set the number of heads N as 12 for our multi-head attention of the transformer encoders in H_{inter} and H_{intra}. The dimension d_n of each head is set as 256. The output dimension of E, H_{inter}, and H_{intra} are 2048. These two attention modules are random initialized. The learning rate is set as $1e^{-4}$ with Adam optimizer in all of our experiments. The batch size is the same as k which is set as 32 for traning DukeMTMC and MARS, yet is variant to window size for ConstructSite ($k \leq 20$) and WILDTRACK ($k \leq 30$), respectively.

5.3. Evaluation settings

We evaluate our model on two experimental settings: video-based Re-ID and MTMC tracking. As we focus only on the tracking algorithm, we use the ground-truth detection bounding boxes for both Re-ID and MTMC tracking. More details are presented as follows.

Video-based Re-ID. During evaluation, we test the model using ground-truth tracklets without IDs across cameras as inputs. That being said, there is no need to define window size (W) in this setting as we test the entire tracklet for each

Table 3: **Comparison of MTMC tracking on the ConstructSite.** The default window size for DeepSort [59] is set as 30.

Method	ConstructSite		
	IDF1	IDP	IDR
GT tracklets+ ResNet-50 [18]	66.50	65.42	66.71
GT tracklets+ NVAN [28]	84.72	87.15	82.63
GT tracklets+ VKD [33]	85.20	84.74	86.91
GT tracklets+ AP3D [16]	84.48	83.64	85.34
GT tracklets+ Ours	**92.38**	**91.31**	**93.47**
DeepSort [59] + ResNet-50 [18]	30.05	21.84	40.16
DeepSort [59] + NVAN [28]	49.16	40.01	56.58
DeepSort [59] + VKD [33]	47.35	36.48	51.31
DeepSort [59] + AP3D [16]	47.56	38.04	53.50
DeepSort [59] + Ours	**62.69**	**61.97**	**63.44**

Table 4: **Ablation studies on the attention modules for video-based Re-ID.** The experiments are conducted on WILDTRACK.

Method	WILDTRACK		
	Rank1	Rank5	mAP
Ours	**85.5**	**96.8**	**72.0**
Ours w/o pos. embeddings	84.2	96.3	71.4
Ours shuffling T sampled images	83.9	96.5	71.5
Ours w/o H_{intra}	82.7	94.1	69.4
Ours w/o H_{inter}	78.5	90.2	67.3

identity in each camera. k denotes the entire tracklets in the testing set or each video, *i.e.*, 625 for MRS and 1110 for DukeMTMC-VideoReID. We employ the standard metrics as in most video-based person Re-ID literature, which are the cumulative matching curve (CMC) used for generating ranking accuracy, and the mean Average Precision (mAP). We report rank-1, rank-5 accuracy and mean average precision (mAP) for evaluation.

MTMC tracking. For MTMC tracking, we first use the single-camera tracking methods with the default window size on the bounding boxes to derive the candidate tracklets for each camera. The number of all tracklets k in the given time window depends on W. Then, we apply the Re-ID model to associate them and across cameras. On the other hand, there is no W for using the ground-truth candidate tracklets. We use ID measures of performance [36] which indicate how well a tracker identifies where the target is. IDP (IDR) is the fraction of computed (true) detections that are correctly identified. IDF1 is the ratio of correctly identified detections over the average number of true and computed detections. ID measures first compute a 1-1 mapping between true and computed identities that maximizes true positives and then compute the ID scores.

Table 5: **Ablation studies on the attention modules for MTMC tracking (GT tracklets).** The experiments are conducted on ConstructSite.

Method	ConstructSite		
	IDF1	IDP	IDR
Ours	**92.38**	**91.31**	**93.47**
Ours w/o pos. embeddings	90.21	91.25	92.41
Ours shuffling T sampled images	90.37	91.12	91.53
Ours w/o H_{intra}	88.77	88.12	89.35
Ours w/o H_{inter}	80.49	80.26	80.25

5.4. Results and comparisons

Re-ID. We compare our Re-ID model with one baseline method (ResNet-50 [18]) and nine state-of-the-art video-based Re-ID approaches, including ETAP-Net [61], STA [13], GLTR [21], TKP [17], COSAM [44], NVAN [28], VKD [33], AP3D [16]. We evaluate our model and these methods on the four datasets and the results is presented in the Table 2. Yet, for the evaluation on the ConstructSite and WILDTRACK, we only run the experiments for those methods whose codes are available. From the table, several phenomenons can be observed which we summarized as two folds. Firstly, our model achieves the best Re-ID performance on ConstructSite and WILDTRACK, which demonstrates that our introduced inter-tracklet and intra-tracklet attention modules are helpful to Re-ID with overlapping cameras. Second, our model exhibits comparable performance with state-of-the-arts on the other two non-overlapping datasets.

MTMC tracking. To apply our Re-ID model for MTMC tracking, we integrate our Re-ID model with the common single-camera tracking approach, i.e, DeepSort [2]. To better analyze the performance exclusively for Re-ID models, we conduct the experiments using ground-truth (GT) tracklets for each camera. This allows us to exclude the errors coming from the single-camera tracking. We also compare our methods with single baseline (ResNet-50) and three of the state-of-the-art Re-ID approaches, which include NVAN [28], VKD [33], AP3D [16]. The results of MTMC tracking is presented in Table 3. Several phenomenons can also be observed. First, our model achieves the best result on both settings: with GT tracklets and with DeepSort [2], which also confirms the effectiveness of our Re-ID model for MTMC tracking. Second, models with DeepSort [2] exhibit inferior tracking performance. This is due to the reason that the single-camera tracking usually generates fragments or leads to several ID switches within single camera.

5.5. Ablation studies

Attention modules. To further analyze the importance of each introduced attention modules which include H_{inter}

Table 6: **Ablation studies of on the number of heads in each module for video-based Re-ID.**

Method	WILDTRACK		
	Rank1	Rank5	mAP
H_{intra}: 24, H_{intra}: 24	85.4	**96.9**	**72.2**
H_{intra}: 12, H_{intra}: 12 (default)	**85.5**	96.8	72.0
H_{intra}: 6, H_{intra}: 12	84.9	96.3	69.7
H_{intra}: 12, H_{intra}: 6	85.2	96.4	71.0
H_{intra}: 6, H_{intra}: 6	79.3	95.4	64.5
H_{intra}: 1, H_{intra}: 1	75.7	91.9	62.7

Table 7: **Ablation studies on the sampling factor T for video-based Re-ID.** The experiments are conducted on WILDTRACK. Note that increasing T will lead to more computational cost.

Method	WILDTRACK		
	Rank1	Rank5	mAP
Ours: $T = 1$	77.4	88.5	63.0
Ours: $T = 4$	84.7	94.2	70.5
Ours: $T = 8$ (default)	**85.5**	96.8	**72.0**
Ours: $T = 12$	85.1	**97.2**	71.5
Ours: $T = 16$	85.2	97.0	71.4

and H_{intra}, we conduct an ablation study shown in Table 4 and Table 5 for Re-ID and MTMC tracking, respectively. Firstly, the intra-tracklet multi-head attention module H_{intra} is shown to be vital to our model since we observe 3% drops at Rank 1 on WILDTRACK and 4% drops at IDF1 on our ConstructSite when the module was excluded. This is caused by no module to learn the temporal relationship within the tracklet. Thus, we can not embed temporal positional patterns into the most representative feature for each tracklet. Secondly, without the inter-tracklet multi-head attention module H_{inter}, our model would not be able to learn discriminative features to perform cross-camera association. This results in a large performance drop (about 7% at Rank1 on WILDTRACK and 10% at IDF1 on ConstructSite).

Number of heads in multi-head attention. We present the performance of our multi-head attention with varying numbers of heads in Table 6. From this table, we see that while such hyperparameters need to be determined in advance, the results were not sensitive to their choices. In other words, with a sufficient number of heads, the model will be able to have satisfactory performance. On the other hand, the model consisting of only one self-attention (as previous attention methods do for Re-ID) in each of the multi-attention modules is unable to learn with multiple various features.

Position embeddings and time synchronization. As shown in Table 4 and Table 5 again, we also conduct another ablation studies on the position embeddings and the timing of the sampled T images for each tracklet. We can observe a slight performance drop when the position embeddings are removed. This infers that the feature of relevant timing position in each tracklet is important to the model. In addition, shuffling the sampled images in each tracklet would also lead to a similar performance drop due to the erasing of the timing information.

Table 8: **Ablation studies of on the window size for MTMC tracking.** Note that $W = 30$ is the default hyperparameter.

Method	ConstructSite		
	IDF1	IDP	IDR
DeepSort [59]+ours: $W = 15$	62.35	61.31	63.41
DeepSort [59]+ours: $W = 30$	**62.69**	**61.97**	**63.44**
DeepSort [59]+ours: $W = 60$	52.49	54.63	56.50
DeepSort [59]+ours: $W = 120$	26.20	25.83	28.12
DeepSort [59]+ours: $W = 360$	18.65	23.44	17.73

Hyperparameters. We now further discuss the design of our model. First, we present the ablation studies on the sampling factor T in Table 7. We can observe that a different number of sampling frames T will have impacts on the model performance. Yet, to balance and computational cost and the performance (following [23, 52]), we select $T = 8$ as our default hyperparameter. Second, we present the ablation studies on the window size W for single-camera tracker (DeepSort [59]).) in MTMC tracking in Table 8. Since the tracker will exhibit errors such as fragmentations and ID-switch when tracking multiple objects in a single camera, a larger window size will lead to more such errors. Thus, the performance of MTMC tracking using Re-ID model will be affected accordingly.

6. Conclusion

In this paper, we proposed a video-based Re-ID model via transformer for MTMC tracking with overlapping cameras. We introduced inter-tracklet (person to person) attention module to learn the correlation between tracklets across multiple views. Also, we introduced another intra-tracklet (person specific) attention module for learning the representative feature for motion and appearance sequence in each tracklet. The experiments on our ConstructSite and WILDTRACK confirmed the effectiveness of our model for Re-ID and uncalibrated MTMC tracking with overlapping cameras. In addition, our model also successfully handled generic Re-ID with non-overlapping cameras, which was confirmed by the experiments on two benchmark datasets.

Acknowledgement: We thank SHIMIZU CORPORATION for the sponsorship and data collection.

References

[1] Maryam Babaee, Ali Athar, and Gerhard Rigoll. Multiple people tracking using hierarchical deep tracklet re-identification. *arXiv preprint arXiv:1811.04091*, 2018. 2

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process.*, 2016. 2, 7

[3] Xiaobin Chang, Timothy M Hospedales, and Tao Xiang. Multi-level factorisation net for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[4] Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3, 6

[5] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[6] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[7] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[8] Qi Chu, Wanli Ouyang, Hongsheng Li, Xiaogang Wang, Bin Liu, and Nenghai Yu. Online multi-object tracking using cnn-based single object tracker with spatial-temporal attention mechanism. In *Int. Conf. Comput. Vis.*, 2017. 2

[9] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[10] Afshin Dehghan, Shayan Modiri Assari, and Mubarak Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[12] Francois Fleuret, Jerome Berclaz, Richard Lengagne, and Pascal Fua. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007. 3

[13] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2, 6, 7

[14] Weihao Gan, Shuo Wang, Xuejing Lei, Ming-Sui Lee, and C-C Jay Kuo. Online cnn-based multiple object tracking with enhanced model updates and identity association. *Signal Processing: Image Communication*, 2018. 2

[15] Xu Gao and Tingting Jiang. Osmo: Online specific models for occlusion in multiple object tracking under surveillance scene. In *ACM Int. Conf. Multimedia*, 2018. 2

[16] Xinqian Gu, Hong Chang, Bingpeng Ma, Hongkai Zhang, and Xilin Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *Eur. Conf. Comput. Vis.*, 2020. 6, 7

[17] Xinqian Gu, Bingpeng Ma, Hong Chang, Shiguang Shan, and Xilin Chen. Temporal knowledge propagation for image-to-video person re-identification. In *Int. Conf. Comput. Vis.*, 2019. 6, 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 6, 7

[19] Mahdi M Kalayeh, Emrah Basaran, Muhittin Gökmen, Mustafa E Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[20] H. W. Kuhn and Bryn Yaw. The hungarian method for the assignment problem. *Naval Res. Logist. Quart*, 1955. 4

[21] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *Int. Conf. Comput. Vis.*, 2019. 6, 7

[22] Peng Li, Jiabin Zhang, Zheng Zhu, Yanwei Li, Lu Jiang, and Guan Huang. State-aware re-identification feature for multi-target multi-camera tracking. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2019. 1, 2

[23] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2, 4, 6, 8

[24] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014. 2, 3

[25] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[26] Yu-Jhe Li, Xinshuo Weng, and Kris Kitani. Learning Shape Representations for Person Re-Identification under Clothing Chang. *WACV*, 2021. 2

[27] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, and Yi Yang. Improving person re-identification by attribute and identity learning. In *arXiv preprint*, 2017. 2

[28] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In *British Machine Vision Conference*, 2019. 6, 7

[29] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[30] Wenqian Liu, Octavia Camps, and Mario Sznaier. Multi-camera multi-object tracking. *arXiv preprint arXiv:1709.07065*, 2017. 2

[31] Yu Liu, Junjie Yan, and Wanli Ouyang. Quality aware network for set to set recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[32] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based per-

son re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 2

[33] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *Eur. Conf. Comput. Vis.*, 2020. 6, 7

[34] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 5

[35] 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 workshop on Benchmarking Multi-Target Tracking*, 2016. 2, 3, 6

[36] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Eur. Conf. Comput. Vis.*, 2016. 7

[37] Ergys Ristani and Carlo Tomasi. Features for multi-target multi-camera tracking and re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 1, 2

[38] Han Shen, Lichao Huang, Chang Huang, and Wei Xu. Tracklet association tracker: An end-to-end learning-based association approach for multi-object tracking. *arXiv preprint arXiv:1808.01562*, 2018. 2

[39] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[40] Hao Sheng, Jiahui Chen, Yang Zhang, Wei Ke, Zhang Xiong, and Jingyi Yu. Iterative multiple hypothesis tracking with tracklet-level association. *IEEE Trans. Circuit Syst. Video Technol.*, 2018. 2

[41] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[42] Francesco Solera, Simone Calderara, and Rita Cucchiara. Learning to divide and conquer for online multi-target tracking. In *Int. Conf. Comput. Vis.*, 2015. 2

[43] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[44] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Int. Conf. Comput. Vis.*, 2019. 6, 7

[45] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal S Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019. 2

[46] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Eur. Conf. Comput. Vis.*, 2018. 2

[47] Siyu Tang, Bjoern Andres, Miykhaylo Andriluka, and Bernt Schiele. Subgraph decomposition for multi-target tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[48] Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Multi-person tracking by multicut and deep matching. In *Eur. Conf. Comput. Vis.*, 2016. 2

[49] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 1, 4

[51] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet. In *ACM Int. Conf. Multimedia*, 2019. 2

[52] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Eur. Conf. Comput. Vis.*, 2016. 4, 6, 8

[53] Yongxin Wang, Kris Kitani, and Xinshuo Weng. Joint Object Detection and Multi-Object Tracking with Graph Neural Networks. *ICRA*, 2021. 2

[54] Yan Wang, Lequn Wang, Yurong You, Xu Zou, Vincent Chen, Serena Li, Gao Huang, Bharath Hariharan, and Kilian Q Weinberger. Resource aware person re-identification across multiple resolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[55] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2, 3

[56] Longyin Wen, Dawei Du, Shengkun Li, Xiao Bian, and Siwei Lyu. Learning non-uniform hypergraph for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 2

[57] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. *IROS*, 2020. 2

[58] Xinshuo Weng, Ye Yuan, and Kris Kitani. PTP: Parallelized Tracking and Prediction with Graph Neural Networks and Diversity Sampling. *Robotics and Automation Letters*, 2021. 2

[59] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE Int. Conf. Image Process.*, 2017. 2, 7, 8

[60] Chih-Wei Wu, Meng-Ting Zhong, Yu Tsao, Shao-Wen Yang, Yen-Kuang Chen, and Shao-Yi Chien. Track-clustering error evaluation for track-based multi-camera tracking system employing human re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2017. 2

[61] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 3, 6, 7

[62] Yu Xiang, Alexandre Alahi, and Silvio Savarese. Learning to track: Online multi-object tracking by decision making. In *Int. Conf. Comput. Vis.*, 2015. 2

[63] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua

Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2015. 1, 4

[64] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2

[65] Yuanlu Xu, Xiaobai Liu, Yang Liu, and Song-Chun Zhu. Multi-view people tracking via hierarchical trajectory composition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016. 3

[66] Kwangjin Yoon, Young-min Song, and Moongu Jeon. Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. *IET Image Processing*, 2018. 2

[67] Fengwei Yu, Wenbo Li, Quanquan Li, Yu Liu, Xiaohua Shi, and Junjie Yan. Poi: Multiple object tracking with high performance detection and appearance feature. In *Eur. Conf. Comput. Vis.*, 2016. 2

[68] 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. 2

[69] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *Eur. Conf. Comput. Vis.*, 2016. 2, 3, 6

[70] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2, 3

[71] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. In *arXiv preprint*, 2016. 1

[72] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Int. Conf. Comput. Vis.*, 2017. 2, 3

[73] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2