Video-based Person Re-identification without Bells and Whistles

Chih-Ting Liu[1], Jun-Cheng Chen[2], Chu-Song Chen[3], Shao-Yi Chien[1]

[1]Graduate Institute of Electronics and Engineering, National Taiwan University

[2]Research Center for Information Technology Innovation, Academia Sinica

[3]Department of Computer Science and Information Engineering, National Taiwan University

jackieliu@media.ee.ntu.edu.tw, pullpull@citi.sinica.edu.tw, chusong@csie.ntu.edu.tw, sychien@ntu.edu.tw

Abstract

Video-based person re-identification (Re-ID) aims at matching the video tracklets with cropped video frames for identifying the pedestrians under different cameras. However, there exists severe spatial and temporal misalignment for those cropped tracklets due to the imperfect detection and tracking results generated with obsolete methods. To address this issue, we present a simple re-Detect and Link (DL) module which can effectively reduce those unexpected noise through applying the deep learning-based detection and tracking on the cropped tracklets. Furthermore, we introduce an improved model called Coarse-to-Fine Axial-Attention Network (CF-AAN). Based on the typical Non-local Network, we replace the non-local module with three 1-D position-sensitive axial attentions, in addition to our proposed coarse-to-fine structure. With the developed CF-AAN, compared to the original non-local operation, we can not only significantly reduce the computation cost but also obtain the state-of-the-art performance (91.3% in rank-1 and 86.5% in mAP) on the large-scale MARS dataset. Meanwhile, by simply adopting our DL module for data alignment, to our surprise, several baseline models can achieve better or comparable results with the current state-of-the-arts. Besides, we discover the errors not only for the identity labels of tracklets but also for the evaluation protocol for the test data of MARS. We hope that our work can help the community for the further development of invariant representation without the hassle of the spatial and temporal alignment and dataset noise. The code, corrected labels, evaluation protocol, and the aligned data will be available at https://github.com/jackie840129/CF-AAN.

1. Introduction

Person re-identification (Re-ID) aims to solve the problem of identifying pedestrians in a multi-camera surveillance system. Many researches focus on the image-based setting that identifies people with still images [19, 23, 36, 50, 46]. Recently, video-based Re-ID [10, 14, 15, 26] has drawn significant attention since comparing continuous

Figure 1: **Video tracklet processed with our re-Detect and Link (DL) module.** (a) The tracklet is with unexpected noise, where the frame at $t = 4$ is dominated by the man in blue shirt, but the ground truth identity is the girl in white dress. (b) The tracklet after DL is less interfered by the man.

video sequences is more practical for the real-world scenarios. Besides, the appearance information with spatial and temporal relations in a video tracklet contains more cues for matching people under different views. The most commonly used methods for tackling video sequences are the 3D convolution [37] and Non-local operation [40], which can effectively aggregate the features along the spatial and temporal dimensions. However, in contrast to image-based setting that the training and testing images of pedestrians are chosen with the least noise from their belonged tracklets, the video-based Re-ID faces more unexpected challenges owing to the imperfect bounding box detection.

MARS [47], the largest video-based Re-ID dataset so far, adopted traditional DPM [8] as the pedestrian detector and applied GMMCP tracker [6] with color histogram as image features, which is not robust enough for linking people under a complicated environment with occlusion. As Fig. 1(a) illustrated, the bounding boxes generated by the weak detector cannot well fit the desired identity (the girl with white dress). Recently, Gu *et al.* [10] proposed the

appearance preserving module (APM) inserted before the 3D convolution to align the features along the temporal axis based on each anchor (the center) frame of the 3D sliding windows. Although the method achieves the state-of-the-art performance, it still cannot resolve the problems when the center frame contains unexpected noise, such as the fourth frame in Fig 1(a), where the APM will align the third and fifth frames (if the filter size along the temporal axis is 3) according to the appearance of the man with blue T-shirt.

Since efficient deep-learning algorithms are well-developed for object detection and tracking in the past few years [34, 27, 2, 41], to help the community for the further development of invariant representation without the hassle of the spatial and temporal alignment, we revised the original dataset with our proposed simple but effective **re-Detect and Link** (**DL**) module. Because we cannot obtain the original video stream containing the whole image frame, our DL module serves as a pre-processing technique on the Re-ID data. Given the original noisy cropped sequence, we first apply a pretrained efficient object detector [2] to generate much tighter bounding boxes. If there are multiple pedestrian candidates, we will link the pedestrians based on their image features using ID-discriminative embedding (IDE) [49]. Last, according to the aspect ratio and the position of the bounding box, we resize and pad it to the desired image size, as shown in Fig 1(b). Surprisingly, with only the input data processed by our DL module first, even the C2D baseline method [10], which only averages the features of each image generated by 2D ResNet [11], or the normal 3D convolution model P3D-C [32] can achieve promising results. As shown in Table 1, we conduct more experiments on the original and the processed data using some recent state-of-the-arts reproduced by ourselves. From the table, it can be seen that originally the AP3D with the APM module proposed by Gu *et al.* [10] (the last row) can boost about 2% in mAP compared to its P3D-C counterpart (the second row). However, with the aligned input images generated by our DL module, it only increase 0.4% in mAP. This shows that the state-of-the-art AP3D cannot extract more discriminative features for Re-ID given the already aligned data. Furthermore, we can see that the self-attention based Non-local Network [10, 26] combined with our DL module can achieve the new state-of-the-arts, which means the self-attention on the less noisy data can generate more representative Re-ID features. Thus, in the next step, we focus on the Non-local Network but developing an efficient baseline model which can perform comparable results.

Non-local Network achieves state-of-the-art performance on video-based Re-ID, but its high computation cost remains an issue for practical usage. Each feature point along spatial and temporal dimensions needs to compute its self-attention map for all other points. To reduce the computation while retaining the performance of

Table 1: Performance of recent state-of-the-arts reproduced with our re-Detect and Link (DL) on MARS [47]. The score with underline is the runner-up.

Method	Original Results		w/ our DL	
	mAP	rank-1	mAP	rank-1
FT-WFT [31]	82.9	88.6	83.8	90.0
P3D-C [32, 10]	83.1	88.5	85.0	91.0
C2D [10]	83.4	88.9	84.9	91.0
Non-Local [10, 26]	85.0	89.6	**86.2**	**91.4**
TCLNet [14]	<u>85.1</u>	<u>89.8</u>	<u>85.8</u>	90.8
AP3D [10]	**85.1**	**90.1**	85.4	<u>91.0</u>

(a) Labeling Error (b) Appears in distractor (c) Ambiguous identity

0142C2T0008 0184C5T0025   0404C2T0015 0000C2T1600   0318C6T0027

Figure 2: Illustration of the labeling errors and ambiguous cases in MARS [47] testing set. More samples and details can be found in Sec. 4.5.

Non-local Network on Re-ID, following the idea of axial-attention [13, 16] and the multi-granularity (coarse-to-fine) structure in [45], we propose the *Coarse-to-Fine Axial-Attention Network* (CF-AAN). With the axial attention, we can factorize the 3D attention operation into three 1-D attention ones sequentially along the height-, width- and temporal-axis. To further boost the efficiency, in contrast to [45] that adding the coarse-to-fine module after the whole model backbone, we directly integrate it into our axial-attention. We split the input tensor into multiple scales along the channel dimension, and transform the spatial dimension from coarse to fine scales. To the best of our knowledge, we are the first to adopt axial-attention in video-based Re-ID. Our DL+CF-AAN approach not only achieves the state-of-the-art performance on two large-scale datasets [47, 42], but also significantly save the computation as compared with vanilla Non-local Network, which can be regarded as an efficient baseline self-attention method.

In addition to the application of our DL module that can significantly improve the performance, we also find that there are multiple **labeling errors or noises** in the MARS testing data. As shown in Fig. 2(a), the two tracklets are labeled as different identities (ID 142 and 184) but are actually the same person. Or in Fig. 2(b), the tracklet with ID 404 in camera 2 also appears in the distractor class (ID 0), which will make the model easily match the two tracklets but counted as an error matching in the evaluation. There are also some ambiguous cases that cannot be distinguished even by human. As in Fig 2(c), the ID 318 is the man in blue behind but the bounding boxes also contains the woman in white (ID 322). Thus, we revise the labels in the testing set and the original evaluation protocol. The details will be described in Sec. 4.5. We hope that the release of our DL

processed test data on MARS can help the community to validate their methods on a clean testing set and push the further development of improved representation.

Our contributions can be highlighted as follows:

- We propose a re-Detect and Link module that can align the noisy tracklet on the image level, which makes a simple method achieving comparable performance.

- Besides the aligned data, we additionally provide revised identity labels and evaluation protocol in MARS testing set, which helps validate the new methods on a corrected benchmark.

- A baseline Coarse-to-Fine Axial Attention Network (CF-ANN) is proposed, which performs axial-attention from coarse to fine levels, which not only reduces the computation cost but achieves the promising performance.

2. Related Work

We briefly review the recent related development of video-based person re-identification as follows.

Video-based Person Re-identification Compared to image-based Re-ID, video-based setting contains more frames and additional temporal information. Typically, researchers aggregate the information among a tracklet with temporal modeling or attention-based algorithms and optimize the model with discriminative learning [49] or metric learning [12]. For temporal modeling, McLaughlin *et al.* [30] first applied Recurrent Neural Network (RNN) on the frame-wise CNN features to allow information flowing among different frames and obtain a sequence-level representation. Inspired by the success of 3D Convolutional Neural Network on action recognition [4, 18], the work [21] first adopted the 3D convolution to automatically learn the relation from low- to high-level features along spatial and temporal dimensions. In order to resolve the alignment problems, Gu *et al.* [10] then proposed an APM module inserted before the 3D convolution to align the features among each 3D filters. In contrast to treating each frame even, some works utilized the attention mechanism that can focus on some specific regions representing the identity better [22, 35, 5, 9, 26, 45]. Li *et al.* [22] proposed multiple spatial attention modules that can focus on many important spatial regions across different frames and the spatial features are then aggregated by a learnable temporal attention. Chen *et al.* [5] adopted a novel co-attention mechanism that can dynamically learn the feature representation based on the query and gallery pairs. Zhang *et al.* [45] explored the attention mechanism with a global reference, which can effectively learn the attention more on the region with close relation to the global guidance. Besides performing attention on the last layer of CNN features, Liu *et al.* [26] started to aggregate the popular non-local self-attention [40] inside the CNN backbone. Compared to those methods, our model

is based on the self-attention operation and added with computation efficient structures into the model design.

Self-Attention Since the self-attention based Transformer [38] obtained a great success in nature language processing, recently many works started to tackle the problems in computer vision with self-attention [40, 1, 7, 33, 13, 39, 3, 16]. The plain type of the self-attention is the non-local network [40] without the position encoding and multi-head attention and was proposed to solve the problem of video classification. Because the non-local self-attention is computation demanding, axial-attention [13, 39] were proposed to factorize the operation into multiple 1-D self-attentions, which can extremely reduce the cost. Dosovitskiy *et al.* [7] and Carion *et al.* [3] even integrated the whole transformer respectively into the image classification and object detection tasks, and they all obtained comparable performance to the methods with original CNN backbone. Our work focus on adding the efficient axial-attention module with our proposed coarse-to-fine structure into typical CNN to learn spatially and temporally attentive feature representation.

Dataset and Evaluation Protocol Revision In the field of person re-identification, there is no work exploring and revising the original imperfect data or discussing the evaluation protocols, labeling errors, and the ambiguous cases in the testing set. We found that in the field of face detection, there are some works investigating the noise in the labels or the bias in evaluation protocols [29, 25, 44]. Mathisas *et al.* [29] provided improved annotations of existing face datasets and evaluation criteria that resolved the original problems. Besides, they also showed that when properly used, a simple vanilla baseline can reach top performance on face detection. Lin *et al.* [25] and Zhang *et al.* [44] both tried to remove the data with labeling errors before training by utilizing the inherent data distributions. Compared to our work, we adopt pretrained deep learning-based object detector to refine the original test data that are unfit to the target identity. With the aligned data, even a simple baseline method can achieve outstanding performance. Moreover, we manually check the errors with the existing Re-ID evaluation protocol and provide some revision of not only the labels but the evaluation protocol.

3. The Proposed Method

Fig. 3 demonstrates the pipeline of our re-Detect and Link (DL) techniques and the proposed Coarse-to-Fine Axial Attention Network (CF-AAN). Given an original imperfect video tracklet $\mathcal{V}$ with N images, $\mathcal{V} = \{I_1, I_2, .., I_N\}$, we first adopt our DL module to obtain the processed tracklet $\mathcal{V}'$, which is more robust and aligned. The detail of our DL will be described in Sec. 3.1. Then, as the typical pipeline of video-based Re-ID, we sample T frames from $\mathcal{V}'$ as the input of our CF-AAN. Our network consists of a backbone CNN and multiple Coarse-to-Fine Axial Attention (CF-AA) modules, which are separately inserted
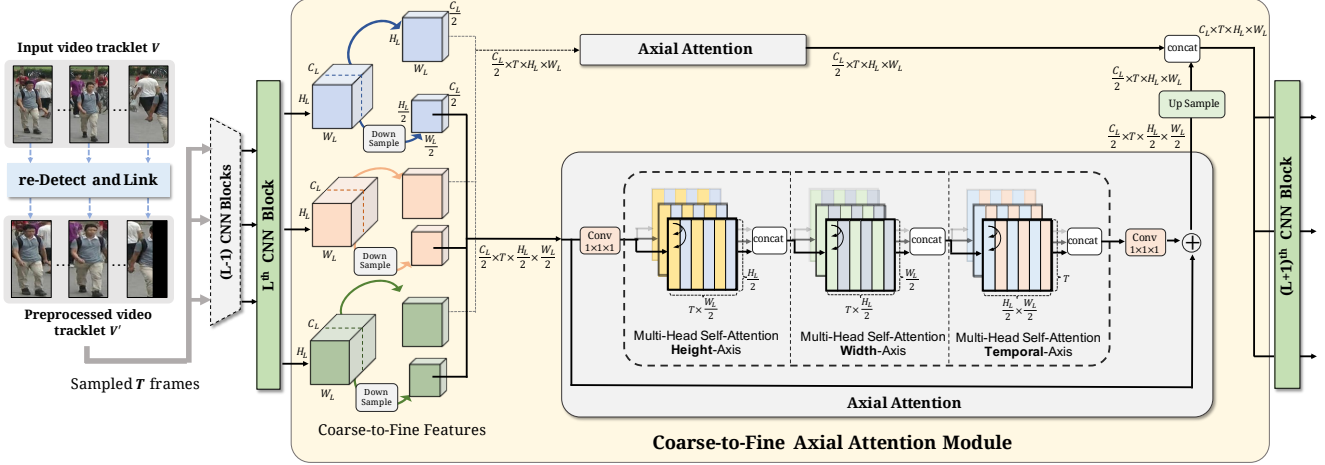
Figure 3: **Pipeline of our DL and CF-AAN architecture.** The original tracklet $\mathcal{V}$ is first fed into the DL module and become the processed tracklet $\mathcal{V}'$, which will then be sampled and fed to CF-AAN. We demonstrate one CF-AA module between the Lth and $(L+1)^{th}$ CNN block. There are two scales of features and the axial-attention will perform on each of them. The outputs will be up-sampled and concatenated to become the input of the next CNN block.

between the CNN blocks. The operations in our CF-AAN are described in Sec. 3.2. Last, in Sec. 3.3, the video features generated by our CF-AAN will be aggregated with the masks created along with DL module and optimized with the common losses for Re-ID.

3.1. Data Alignment with re-Detect and Link

With the noisy video tracklet $\mathcal{V}$ with N images, we sequentially perform our re-Detect and Link (DL) method on each video frame and create a new processed tracklet $\mathcal{V}'$ with N frames, too. As illustrated in Fig. 4, first, all images are padded and fed to the object detector [2] to generate candidate bounding boxes with the "person" class. For the first frame, if there are multiple candidates, we will assume that the bounding box with larger area is the desired one. Then, similar to the feature-based real-time object tracking [41], we extract the feature f of the cropped image I_1' by the IDE feature extractor trained on the original dataset [47], and save it as the global feature $f_g = f_1$. Next, for each consecutive frame i, if there are multiple candidates, we will compare each extracted feature f_i^j to the global feature f_g and choose the one with the smallest Euclidean distance, where j is the index of the candidate bounding box in i^{th} frame. After choosing the candidate for the ith frame, the global feature will then be updated by

$$f_g = \alpha f_g + (1-\alpha) f_i , \qquad (1)$$

where α is set to 0.9 in our case.

Note that in Re-ID datasets, we cannot obtain the original full image frame captured by cameras and perform our DL method. Thus, after we apply object detection on the noisy cropped image, we may obtain a new cropped identity with only part of his/her appearance, as shown in Fig 1(b).
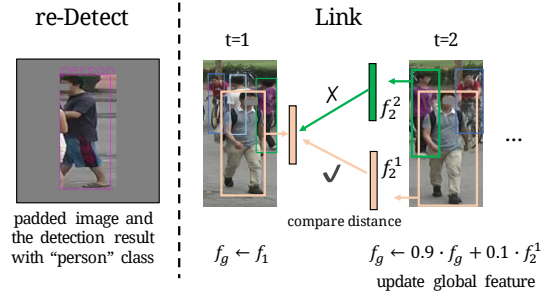

Figure 4: Illustration of the re-Detect and Link module.

According to the aspect ratio and the position of the bounding box in the image, if the bounding box is slim (the height is much larger than the width) and its position is on the left (right) of the image, we will shift it to the right (left), resize it based on its original aspect ratio and pad it to the desired image size. Furthermore, we also create a mask M_i of the output image I_i' representing whether each pixel is the padded one or not. This mask will then be applied in the feature aggregation of our CF-AAN, which will be described in Sec. 3.3.

Discussion Comparing to other methods proposing an automatically learned feature alignment mechanism inside their backbone model [51, 10], our DL module adopts an additional object detector to help reduce the original noise in the data. It seems that our method requires additional computation cost but and utilizes extra information. However, we want to point out that the goal behind our DL module is to simulate a nowadays real-life scenario with efficient and robust deep learning-based object detection and tracking **before** Re-ID. Thus, when it really comes to the Re-ID phase, actually there has been no need for this additional

cost of DL module on the input tracklet. Furthermore, as shown in the Table 1, with the aligned data, the simplest baseline can obtain a promising Re-ID result and the original state-of-the-art methods that specifically deal with the problems of misalignment will not retain its competitiveness. We think that with the release of the data processed by our simple alignment method, it can help the community explore more on the attention-based methods or the methods for learning invariant feature representation.

3.2. Coarse-to-Fine Axial-Attention Network

As shown in Table 1, the existing self-attention based Non-local Network can achieve the best result on the aligned data. However, the efficiency is the main drawback. We propose a simple method called Coarse-to-Fine Axial-Attention Network that contains a coarse-to-fine mechanism and a position-sensitive axial-attention which dramatically reduce the computation burden but retain comparable performance.

Self-Attention: We first introduce the typical 3D self-attention [40] operation as follows. Given an input feature map $x \in \mathbb{R}^{C_{in} \times T \times H \times W}$ with channels C_{in}, temporal length T, height H, and width W, the output y at position $o = (i, j, t)$, $y_o \in \mathbb{R}^{C_{out}}$, is computed by aggregating all the projected input as :

$$y_o = \sum_{p \in \mathcal{N}} \text{softmax}_p(q_o^T k_p) v_p \qquad (2)$$

where $\mathcal{N}$ is the set of the whole HWT locations, and queries q_o, keys k_o, and values v_o are three different linear projections of the input x_o, $\forall o \in \mathcal{N}$ from dimension C_{in} to intermediate $C_{q,k}$ for query and key projection or C_{out} for value projection. As opposed to convolution which only captures local relations, this mechanism allows us to capture related but non-local context in the whole feature map. Commonly, it will be inserted into multiple locations between the backbone CNN layers, and each complexity is $\mathcal{O}(H^2 W^2 T^2)$.

Axial-Attention: To reduce the computation of non-local self-attention, in 2D image classification tasks, the axial-attention has been proposed [13], they factorized the 2D self-attention operation into two 1-D axial-attentions. When applied to our video-based Re-ID, the 3D self-attention will be consecutively factorized into height-axis, width-axis and the temporal-axis. With this transformation, the complexity can be reduced to $\mathcal{O}(H^2 WT + HW^2 T + HWT^2)$. The formulation of the axial-attention, with the height-axis as an example, is as follows.

$$y_o = \sum_{p \in \mathcal{N}_{H \times 1 \times 1}} \text{softmax}_p(q_o^T k_p) v_p \qquad (3)$$

where the location p only lies along the H axis.

Furthermore, based on the concept proposed in the Transformer [38], many works start to encode the positional encoding into the self-attention structure [3, 1, 33]. Thus, the final method we adopt is based on the positional-sensitive axial-attention proposed in [39], where the learnable positional encoding vectors depends on the query vectors, key vectors and the value vectors. The formulation is as follows with the height-axis as an example.

$$y_o = \sum_{p \in \mathcal{N}_{H \times 1 \times 1}} \text{softmax}_p(q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)$$

$$(v_p + r_{p-o}^v) \qquad (4)$$

where the r_{p-o}^q, r_{p-o}^k, and r_{p-o}^v are the learned relative positional embedding. Besides, in practice, as shown in Fig 3, we will extend the single-head attention into multi-head attention to generate a mixture of affinities. To retain the complexity, if there are M parallel single-head attentions, in the mth head, each dimension of the q^m, k^m, and v^m will be shrunk to $\frac{C_{q,v}}{M}$ and $\frac{C_{out}}{M}$. The dimension of the learnable positional vectors r_{p-o}^q, r_{p-o}^k and r_{p-o}^v are also shrunk but the vectors are shared across each head. Thus, the final output z_o will be the concatenation of each head, $z_o = concat_m(y_o^m)$, with the same dimension C_{out}. Last, after conducting the axial-attention (AA) along the three dimensions, we will project the output feature from dimension C_{out} back to C_{in} and added with the input tensor x to become a new refined tensor x', which is formulated as follows.

$$x' = x + Conv(\text{AA}^T(\text{AA}^W(\text{AA}^H(x)))) \qquad (5)$$

Coarse-to-Fine Axial Attention: In addition to multi-head attention that learns different structure of affinities, we propose a Coarse-to-Fine Axial-Attention module (CF-AA) that not only makes the self-attention learn on different scales of the spatial dimension but further reduce the computation. Different from [45], which can only perform multi-scale structure on the last layer of CNN backbone with the smallest resolution, we can apply our structure along with the axial-attention from the mid-level stage to high-level stage inside the backbone. As shown in Fig. 3, we split the input tensor x with S scales along the channel dimension and for the sth scale, we downsample the spatial resolution to $H_s \times W_s$, where $H_s = \frac{H}{2^{s-1}}$ and $W_s = \frac{W}{2^{s-1}}$. Thus, if $S = 2$ as an example, the original input tensor x will be split into $x_1 \in \mathbb{R}^{\frac{C_{in}}{2} \times T \times H \times W}$ with a fine scale and $x_2 \in \mathbb{R}^{\frac{C_{in}}{2} \times T \times \frac{H}{2} \times \frac{W}{2}}$ with a coarse scale. The split tensors are then separately fed into the axial-attention and the outputs are upsampled and concatenated along the channel dimension in order to retain the original tensor size.

3.3. Feature Aggregation and Optimization

Our CF-AAN contains a 2D CNN backbone and several CF-AA modules inserted between the CNN blocks. Af-

ter the last CNN layer, there will be T tensors with size $\mathbb{R}^{C' \times H' \times W'}$. As mentioned in Sec. 3.1, because there are some input pixels which are the padded ones without any information, we first downsample the mask M to M' according to the spatial dimension H' and W', and utilize the mask to average-pool on the desired spatial region to generate T vectors with C' dimension. Then, we aggregate the features with the typical average operation followed by a Batch-Normalization (BN) layer [17] to create the final feature representation $f_\mathcal{V}$ of the video tracklet. To optimize the network, we follow the two loss combinations in BoT [28], which consists of a batch-hard triplet loss [12] on the features before BN and a cross-entropy loss [49] on the identity classifier (a fully-connected layer) after the feature $f_\mathcal{V}$.

4. Experiments

In this Section, we conduct extensive evaluation and ablation studies of the proposed approach in addition to the analysis and correction of data noise and labeling errors for the evaluation dataset.

4.1. Datasets and Evaluation Protocol.

We evaluate the proposed method on two large-scale datasets, MARS [47] and DukeMTMC-VideoReID [42], abbreviated as DukeV. MARS consists of 17,503 tracks and 1,261 identities. Each track has 59 frames on average. Deformable Part Model [8] is employed to detect pedestrians and GMMCP [6] is used to track pedestrians. To make the MARS dataset even more challenging, they include 3,248 distractor tracks. For DukeV, it comprises 4,832 tracks and 1,404 identities. Each track contains 168 frames on average. Different from MARS, the detection and tracking ground truth are manually labeled. We use the rank-1 (R1) in the Cumulative Matching Characteristics (CMC) and the mean Average Precision (mAP) [48] as evaluation metrics.

4.2. Implementation Details.

re-Detect and Link. Our object detector is the YoloV4 [2] pretrained on the COCO dataset [24]. The IDE [47] model for linking the candidates is a ResNet-50 [11]. We perform our DL module both on MARS and DukeV dataset. However, because only the MARS dataset is adopted with traditional detector and tracker, where the data in DukeV is manually labeled, the processed data of DukeV is almost the same as before.

CF-AAN. For our CF-AAN, we adopt ImageNet pretrained ResNet-50 [11] as our backbone. Similar to the structure of Non-local Network [40], we insert 5 CF-AA modules, 2 after $conv3\_3$, $conv3\_4$ and another 3 after $conv4\_4$, $conv4\_5$, and $conv4\_6$ respectively. In our coarse-to-fine structure, we split the feature into four levels ($S = 4$) and in each axial-attention, we set the number of head $M = 2$. Thus, the total number of heads in a coarse-to-fine axial-attention module is equals to 8, which is similar to the original axial-attention network [39]. In

Figure 5: Examples of video tracklets processed by our DL.

the training stage, we sample $T = 6$ images as an input tracklet. Each frame in a tracklet is resized to 256×128 and synchronously augmented with random horizontal flip. As for the optimizer, Adam [20] with weight decay 5×10^{-5} is adopted. We train the model for 220 epochs. The learning rate is initialized to 10^{-4} and multiplied by 0.1 after every 50 epochs. In the testing stage, for each tracklet, we split it into several 6-frame clips, and then the feature representations for each clip are averaged to become the final representation.

4.3. Ablation Study

In Table 2, we conduct ablation study on our proposed re-Detect and Link (DL) module and our Course-to-Fine Axial-Attention Network (CF-AAN). Besides the Re-ID performance, we also calculate the computation cost of inference in terms of GFLOPs. We first analyze the effectiveness of the DL module on our baseline method (the first two rows). Our "Baseline" method, with 24.52 GFLOPs operations, contains the same ResNet-50 backbone, types of losses and training details but without all the axial-attention modules, which is just the average of features in each frame. We can clearly see that with the aligned data processed by DL, there is an obvious improvement of the performance (1.7% in mAP). Thus, the alignment of the input video tracklet is crucial and important for the subsequent feature extraction. We also demonstrate some extra examples in Fig. 5. We can see that the problems of misalignment in the left tracklet and the multiple candidates in the right tracklet are resolved after processed with the DL module.

Next, we compare the self-attention based methods. The first one is Non-local Network (the 3th row), which is with single head 3D self-attention but without the positional encoding. Although it can improve about 1.1% in mAP compared to the baseline, the computation also increases (+**17.213** GFLOPs), which is extremely large and almost equal to the baseline. After replacing the operation with axial-attention, the computation can reduce to only +0.361 GFLOPs, while the performance slightly decrease owing to its factorized self-attentions. With the multi-head structure (the 5th row), it can retain the computation cost but increase the performance.

Table 2: **The Ablation Study of our DL and CF-AAN.** We compare the effectiveness of our DL and all the components in CF-AAN with the computation cost (GFLOPs) and performance on MARS. Except the baseline itself, all other computation costs are the increase comparing to the baseline method. C_B: the computation cost of the baseline method.

| Method | w/ our DL | Self-attention Module | | | | #GFLOPs | MARS | |
		Self-attention	# of heads	Posi. Encoding	# of scales		mAP	R-1
Baseline	✘	✘	✘	✘	✘	24.520 (C_B)	83.4	87.7
	✔	✘	✘	✘	✘		85.1	89.7
Non-local	✔	3D self-attention	1	✘	1	C_B+17.213	86.2	**91.4**
Axial-based	✔	Axial-attention	1	✘	1	C_B+0.361	86.0	91.1
	✔	Axial-attention	8	✘	1	C_B+0.361	86.2	91.2
	✔	Axial-attention	8	Sinusoidal	1	C_B+0.377	86.0	91.1
	✔	Axial-attention	8	Relative	1	C_B+0.424	86.4	91.2
	✔	Axial-attention	8	Relative	2	C_B+0.245	86.4	91.3
	✔	**Axial-attention**	**8**	**Relative**	**4**	C_B**+0.126**	**86.5**	91.3

We then apply two types of positional encoding to explore their effectiveness. The first one is the sinusoidal encoding (the 6th row) which is the same as the experiments in [1] and the learnable relative positional embedding (the 7th row) proposed in [39]. We can see that there is no significant influence of all kinds of positional encoding but the relative and learnable characteristics are the best for Re-ID, which can achieve 86.4% in mAP. Last, in the last two rows, we demonstrate the benefits brought by our coarse-to-fine structure. We can see that, because the spatial dimensions decrease in the coarser scale, the total operations also decrease. When the number of scales is 4, the operation can increases only 0.126 GFLOPs compared with the baseline, which is only about **1**% of those in Non-local Network. Furthermore, owing to the coarse-to-fine structure that makes the self-attention learn on different scales, the performance even increases to **86**.**5**% in mAP on MARS dataset. The CF-AAN with four scales is our final model performing the video-based Re-ID.

4.4. Comparison with State-of-the-art Approaches

We compare recent state-of-the-art approaches with our methods on MARS and DukeV datasets in Table 3. We can see that in the past, the methods that globally perform attention mechanism on the last CNN features are the mainstream for dealing with video tracklet [22, 35, 5, 9]. However, the noise and unaligned appearance between frames make it hard to learn a robust attention score. In another way, TCLNet [14] conduct the attention frame by frame, which is less interfered by the alignment problems. AP3D [10] is the recent work that adopts 3D convolution with a feature alignment module inserted between 3D CNN blocks. We can see that once reducing this unaligned problem, a 3D CNN can achieve the best results (in R-1). The MG-RAFA [45] is also the attention-based method, but they adopt the multi-granularity (multi-scales) structure on the output of the CNN features, where the features will then be fed to their global attention methods. This structure obtains

Table 3: **Comparison with state-of-the-arts** (%)**.** The score with underline is the runner-up.

| Method | MARS | | DukeV | |
	mAP	R-1	mAP	R-1
DRSA (CVPR18)[22]	65.8	82.3	-	-
EUG (CVPR18)[43]	67.4	80.8	78.3	83.6
DuATM (CVPR18)[35]	67.7	81.2	-	-
TKP (ICCV19)[22]	73.3	84.0	91.7	94.0
M3D (AAAI19)[21]	74.1	84.4	-	-
Snippet (CVPR18)[5]	76.1	86.3	-	-
STA (AAAI19)[9]	80.8	86.3	94.9	96.2
VRSTC (CVPR19)[15]	82.3	88.5	93.5	95.0
NVAN (BMVC19)[26]	82.8	90.0	94.9	96.3
FT-WFT (AAAI20)[31]	82.9	88.6	-	-
TCLNet (ECCV20)[14]	85.1	89.8	<u>96.2</u>	**96.9**
AP3D (ECCV20)[10]	85.1	<u>90.1</u>	95.6	96.3
MG-RAFA (CVPR20)[45]	<u>85.9</u>	88.8	-	-
DL+CF-AAN (**Ours**)	**86.5**	**91.3**	96.2	<u>96.7</u>

the best results in mAP. Our method consists of a simple but effective pre-processing DL module followed by an extremely efficient CF-AAN. Different from [45], our coarse-to-fine structure is inserted with the axial-attention module between the backbone CNN blocks. We can see that our methods achieve promising performance, which outperform AP3D [10] **1**.**4**% in mAP and **1**.**2**% in R-1 on the MARS dataset. Although the data in DukeV are manually labeled, our model still can retain comparable performance. Thus, in summary, with almost no extra computation cost compared to the baseline, where conducting the DL module is also effortless in real-life scenario, we are the state-of-the-art in terms of the popular mAP metric for the video-based person Re-ID task.

4.5. Label Cleaning and New Evaluation Protocols

As described in Sec. 1, we found some labeling errors or ambiguous cases in the MARS dataset. Thus, we manually check the testing data of the unmatched ones in evaluation and propose a new protocol which additionally address three kinds of new situations: labeling errors, duplication in distractor, and ambiguous identity.
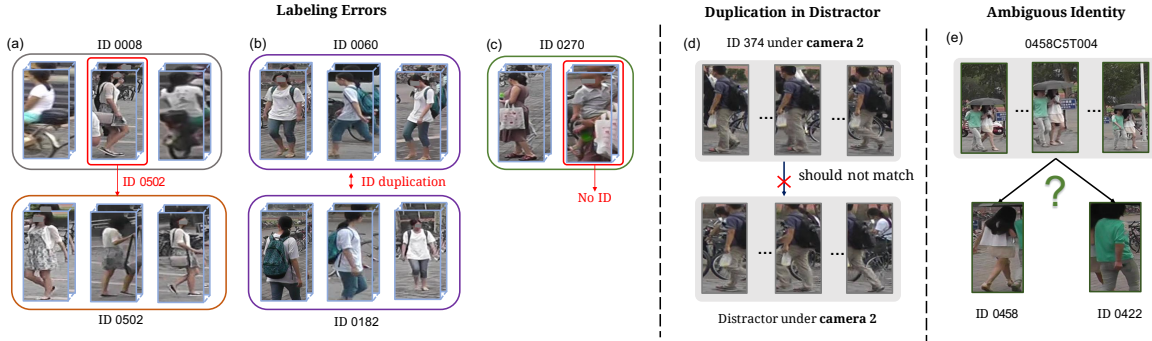
Figure 6: Three kinds of label noises in the MARS testing data.

Purely labeling errors by annotators: There are also three kinds of labeling errors shown in Figs. 6(a)-(c). The first one is that a tracklet may be annotated as another existing identity (6(a)). Or, there are completely two groups of tracklets labeled as a different person but in fact with the same identity (6(b)). Sometimes the tracklet does not belong to any other identities in the testing set. As Fig. 6(c) shows, the identity 270 is the woman but the tracklet marked with red box is the baby she holds. For those three cases, we fix the annotation with the correct or new identity.

Duplication in Distractor Class: In the original evaluation protocol of MARS [47], if a query tracklet matches a gallery tracklet with the same identity but under the same camera, this match will be ignored because Re-ID aims at matching pairs across cameras. However, the "distractor class (ID 0)" in MARS consists of not only the false positive bounding boxes created by pedestrian detector but also some duplicated bounding boxes of the tracklets in testing set. As shown in Fig. 6(d), the tracklet with ID 374 under camera 2 will easily match the same tracklet in distractor and strangely counted as an incorrect match. Thus, we revise the evaluation protocol that if a tracklet matches the other one under the same camera with its same identity or the distractor class, they will both be ignored.

Ambiguous Identity: There are some ambiguous cases in the dataset. As the tracklet in Fig. 6(e), the unfit bounding box contains two persons (ID 485 and ID 422) from the beginning to the end of the tracklet. With our DL, there is only one person left but the true identity cannot be even distinguished by human. For those cases, we will add an additional ambiguous identity of the tracklet and in the evaluation process, the matches of those identities will all be counted as the correct ones.

Similar to Table 1, we reproduce some existing methods not only with data processed by our DL but evaluated under our new protocols, which are shown in Table 4. Furthermore, with their released codes, we also demonstrate the computation cost in inference time with fairly 6-frames clip as input data in terms of GFLOPs. We can see that all methods can improve largely by 2.5% in mAP, but our CF-AAN

Table 4: Performance evaluated with/without new evaluation protocols (N.E.) and the computation cost of recent methods with DL on MARS [47].

Method (**w/ our DL**)	w/o N.E. (mAP)	**w/ N.E.** (mAP)	# GFLOPs
C2D [10]	84.9	87.5	**24.520**
P3D-C [32, 10]	85.0	87.5	26.030
AP3D [10]	85.4	88.2	26.369
TCLNet [14]	85.8	88.4	30.150
Non-Local [10, 26]	<u>86.2</u>	<u>88.6</u>	41.733
CF-AAN (**ours**)	**86.5**	**88.9**	<u>24.646</u>

still achieves the best result (88.9% in mAP). When regarding the computation cost, those of our CF-AAN are comparable to the ones of the simplest C2D baseline method and promisingly, also lower than all existing state-of-the-arts.

5. Conclusion

In this work, we present a simple re-Detect and Link module to further process the Re-ID datasets, which can significantly refine the data generated with obsolete methods. Furthermore, the proposed Coarse-to-Fine Axial-Attention Network significantly improves the original non-local module in terms of computational cost with three 1-D position-sensitive axial-attentions and the proposed coarse-to-fine structure while achieving the state-of-the-art performance. With our refined data, we find that several baseline models can achieve comparable results with current state-of-the-arts. In addition, we also disclose the errors not only for the identity labels but also the evaluation protocol for the test data of MARS. With these findings, we hope the release of corrected data can encourage the community for the further development of invariant representation on view, pose, illumination, and other variations without the hassle of the spatial and temporal alignment and dataset noise.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Int. Conf. Comput. Vis.*, pages 3286–3295, 2019.

[2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *Eur. Conf. Comput. Vis.*, 2020.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6299–6308, 2017.

[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.*, pages 1169–1178, 2018.

[6] 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.*, pages 4091–4099, 2015.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–8. IEEE, 2008.

[9] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*, volume 33, pages 8287–8294, 2019.

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

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016.

[12] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[13] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*, 2019.

[14] Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Temporal complementary learning for video person re-identification. *Eur. Conf. Comput. Vis.*, 2020.

[15] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7183–7192, 2019.

[16] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Int. Conf. Comput. Vis.*, pages 603–612, 2019.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, 2012.

[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.*, pages 1062–1071, 2018.

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale 3d convolution network for video based person re-identification. In *AAAI*, volume 33, pages 8618–8625, 2019.

[22] 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.*, pages 369–378, 2018.

[23] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2285–2294, 2018.

[24] 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 *Eur. Conf. Comput. Vis.*, pages 740–755. Springer, 2014.

[25] Wei-An Lin, Jun-Cheng Chen, and Rama Chellappa. A proximity-aware hierarchical clustering of faces. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 294–301. IEEE, 2017.

[26] 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. *Brit. Mach. Vis. Conf.*, 2019.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Eur. Conf. Comput. Vis.*, pages 21–37. Springer, 2016.

[28] 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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[29] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *Eur. Conf. Comput. Vis.*, pages 720–735. Springer, 2014.

[30] Niall McLaughlin, Jesus Martinez Del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1325–1334, 2016.

[31] Priyank Pathak, Amir Erfan Eshratifar, and Michael Gormish. Video person re-id: Fantastic techniques and where to find them, 2019.

[32] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *Int. Conf. Comput. Vis.*, pages 5533–5541, 2017.

[33] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. *arXiv preprint arXiv:1906.05909*, 2019.

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, pages 91–99, 2015.

[35] 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.*, pages 5363–5372, 2018.

[36] 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.*, pages 480–496, 2018.

[37] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Int. Conf. Comput. Vis.*, pages 4489–4497, 2015.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, pages 5998–6008, 2017.

[39] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. *Eur. Conf. Comput. Vis.*, 2020.

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7794–7803, 2018.

[41] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. *arXiv preprint arXiv:1909.12605*, 2019.

[42] 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.*, June 2018.

[43] 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.*, pages 5177–5186, 2018.

[44] Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Global-local gcn: Large-scale label noise cleansing for face recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7731–7740, 2020.

[45] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10407–10416, 2020.

[46] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3186–3195, 2020.

[47] 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.*, pages 868–884. Springer, 2016.

[48] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jing-dong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Int. Conf. Comput. Vis.*, pages 1116–1124, 2015.

[49] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[50] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2138–2147, 2019.

[51] Zhedong Zheng, Liang Zheng, and Yi Yang. Pedestrian alignment network for large-scale person re-identification. *IEEE Trans. Circuit Syst. Video Technol.*, 29(10):3037–3045, 2018.