Temporal Correlation Vision Transformer for Video Person Re-Identification

Pengfei Wu[1], **Le Wang**[1*], **Sanping Zhou**[1], **Gang Hua**[3], **Changyin Sun**[2]
[1]National Key Laboratory of Human-Machine Hybrid Augmented Intelligence,
National Engineering Research Center for Visual Information and Applications,
and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
[2]School of Artificial Intelligence, Anhui University
[3]Wormpex AI Research
xjtustuwpf@stu.xjtu.edu.cn, {lewang,spzhou}@xjtu.edu.cn, ganghua@gmail.com, cysun@ahu.edu.cn

Abstract

Video Person Re-Identification (Re-ID) is a task of retrieving persons from multi-camera surveillance systems. Despite the progress made in leveraging spatio-temporal information in videos, occlusion in dense crowds still hinders further progress. To address this issue, we propose a Temporal Correlation Vision Transformer (TCViT) for video person Re-ID. TCViT consists of a Temporal Correlation Attention (TCA) module and a Learnable Temporal Aggregation (LTA) module. The TCA module is designed to reduce the impact of non-target persons by relative state, while the LTA module is used to aggregate frame-level features based on their completeness. Specifically, TCA is a parameter-free module that first aligns frame-level features to restore semantic coherence in videos and then enhances the features of the target person according to temporal correlation. Additionally, unlike previous methods that treat each frame equally with a pooling layer, LTA introduces a lightweight learnable module to weigh and aggregate frame-level features under the guidance of a classification score. Extensive experiments on four prevalent benchmarks demonstrate that our method achieves state-of-the-art performance in video Re-ID.

Introduction

Video Person Re-Identification (Re-ID) is a key component of surveillance systems (Zheng et al. 2016; Chen et al. 2018a; Wang et al. 2014). It is different from image-based Re-ID as it retrieves persons from video sequences, thus providing extra temporal clues. With the rise of deep learning techniques, there has been considerable progress in video person Re-ID. However, it is still challenging due to the occlusion caused by dense crowds. Therefore, the focus of video person Re-ID research is on how to exploit temporal information without the interference of occlusion.

The existing methods (Liu et al. 2021a; Yang et al. 2020; Yan et al. 2020; He et al. 2021b; Gu et al. 2020) can be divided into two categories. The first is the one-stage method (Liu et al. 2021a; Yang et al. 2020; Yan et al. 2020; He et al. 2021b; Gu et al. 2020), which utilizes 3D convolution or graph neural networks to learn spatial-temporal information from videos. As mentioned in (Wu et al. 2022), 3D convolution-based methods are often affected by misalignment of adjacent frames and the occlusion problem. Furthermore, graph neural networks (Liu et al. 2021a) usually require an additional pose estimation network to model the body relationships of the target person across frames. The second type is the two-stage method (Wu et al. 2022; Hou et al. 2019), which first extracts frame-level features separately and then performs temporal feature aggregation. This kind of method uses attention mechanisms or generative adversarial networks (GANs) (Hou et al. 2019) to reduce the impact of occlusion within frames, but usually treats video frames equally during the aggregation stage.

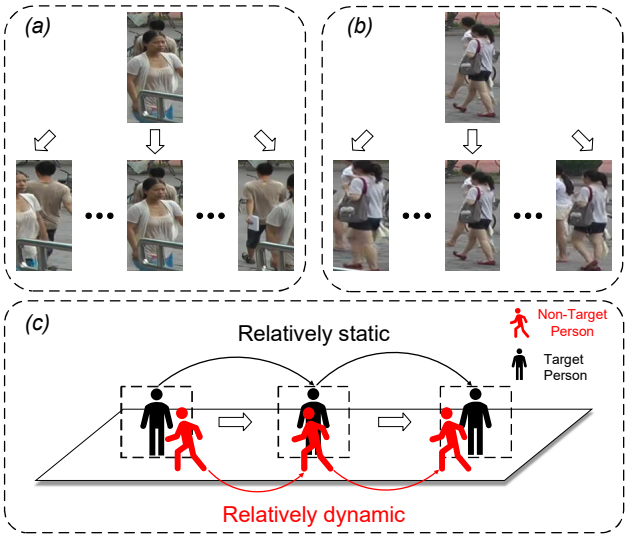

Figure 1: Examples of occlusion cases in video person Re-ID. (a) The target person is the occluded one, while the non-target person approaches from another direction. (b) The target person is the occluder, while the non-target person passes the target at a faster pace. (c) The target person is in the center of the frames and relatively static to the bounding box.

As seen in Figure 1 (c), each frame in videos is cropped according to the bounding box generated by the video person detector (Girshick 2015; Xu, Hrustic, and Vivet 2020), which ensures that the target person is always in the center of

*Corresponding author.

each frame and relatively static to the bounding box. On the other hand, when the non-target person approaches from another direction or passes the target person at a faster pace, the non-target person is relatively dynamic to the bounding box. As illustrated in the first row of Figure 1 (a) and (b), it is difficult to determine whether the target person is an occluder or an occluded one based on a single frame. However, by referring to other frames and considering their relative state, it can be inferred that the target person is the occluded one in (a) and the occluder in (b). Thus, we can use the relative state to mine information about the target person and reduce the impact of most occlusion cases (*e.g.*, the non-target person is relatively dynamic to the bounding box).

Motivated by the above observations, we propose a Temporal Correlation Vision Transformer (TCViT) to tackle the occlusion problem. This two-stage approach first extracts frame-level features using Vision Transformer (ViT) and then enhances the features of the target person based on their relative state. Thereafter, frame-level features are aggregated as the final video representation based on their completeness. To do this, we introduce a parameter-free Temporal Correlation Attention (TCA) module, which aligns the video frames using a kernel correlation filtering algorithm and then boosts the target person's portion in the frame-level features according to temporal correlation. Additionally, unlike other two-stage approaches that treat video frames equally, we employ a lightweight Learnable Temporal Aggregation (LTA) module to weigh and aggregate frame-level features based on classification scores.

We conduct extensive experiments on four prevalent datasets to evaluate our method. The results demonstrate that our method achieves competitive performance with state-of-the-art methods, validating its effectiveness. In summary, our contribution is threefold.

- We propose a Temporal Correlation Vision Transformer (TCViT) for video person Re-ID, which exploits relative state to learn robust features from the target person and aggregate them based on completeness.
- We design a parameter-free Temporal Correlation Attention (TCA) module to solve the occlusion problem, which first aligns the frame-level features by the correlation filter and then re-weights them according to temporal correlation.
- We design a lightweight Learnable Temporal Aggregation (LTA) module to replace the equal treatment strategy, which weighs and aggregates frame-level features under the guidance of classification scores.

Related Work

Video Person Re-ID. Along with the achievement in image-based Re-ID (Chen et al. 2018b; Sun et al. 2018; Zheng et al. 2019; Zhang et al. 2019; Kalayeh et al. 2018), much progress has been made in video-based Re-ID. Existing video Re-ID methods (Bai et al. 2022; Zhou et al. 2017; Aich et al. 2021; McLaughlin, del Rincon, and Miller 2016; Yang et al. 2020; Song et al. 2018) mainly focus on exploiting spatio-temporal clues in videos. Widely used techniques, such as optical flow (McLaughlin, del Rincon, and Miller 2016; Chung, Tahboub, and Delp 2017; Chen et al. 2020), recurrent neural networks (Zhou et al. 2017; McLaughlin, del Rincon, and Miller 2016), graph convolution (Yang et al. 2020; Yan et al. 2020), and 3D convolution (Gu et al. 2020; Li, Zhang, and Huang 2019), are employed to model spatio-temporal relations. However, occlusion and misalignment problems often corrupt the learned features.

Recently, some methods (Gu et al. 2020; Hou et al. 2019, 2020) have been proposed to tackle misalignment and occlusion issues. Gu *et al.* (Gu et al. 2020) reconstruct the feature maps of its adjacent frames to the central frame to ensure feature alignment. Hou *et al.* (Hou et al. 2019) use information from whole frames to restore occluded body parts in occluded frames. However, these approaches can address only one of the misalignment and occlusion problems. Another line of work (He et al. 2021b; Liu et al. 2021a; Yan et al. 2020) exploits the correlation between frames to address occlusion and misalignment implicitly. For example, Liu *et al.* (Liu et al. 2021a) employ a keypoint estimator to extract local features from body parts and interact with corresponding ones across frames. He *et al.* (He et al. 2021b) divide the feature map extracted by CNN into several horizontal parts and then pay dense attention to multi-scale and multi-granularity local features under the guidance of global features. However, the above methods ignore the fixed body structure of humans and the different relative states between the target person and the occlusion. In contrast, we jointly address the problems of occlusion and misalignment and differentiate the target person from occlusion by their different relative states.

Correlation Filtering. Correlation filters are widely explored in object tracking (Henriques et al. 2014, 2012; Bolme et al. 2010; Dai et al. 2019; Ma et al. 2015). Bolme *et al.* (Bolme et al. 2010) learn a minimum output sum of the squared error filter and update it with average moving. Henriques *et al.* (Henriques et al. 2012) augment the training samples by cyclic shift and speed up the algorithm with the kernel method. Based on it, KCF (Henriques et al. 2014) exploits the histogram of oriented gradients (HOG) feature (Dalal and Triggs 2005) to improve the accuracy of the tracker. Zhang *et al.* (Zhang et al. 2014) model the scale change and learn the filters with context information. Danelljan *et al.* (Danelljan et al. 2014) learn an adaptive correlation filter and adopt the color attributes of the target object to object tracking. As for the proposed method, we directly perform correlation filtering on RGB images and use the correlation filter to calculate the deviation of the target person across frames so as to tackle the problem of misalignment efficiently.

Temporal Correlation Vision Transformer

The framework of our proposed TCViT, as shown in Figure 2, consists of TCA and LTA modules. The encoder divides each frame into patches and generates frame-level features. Meanwhile, the TCA module exploits the correlation filtering algorithm to align the patches and then enhances the patches of the target person based on temporal correlation. Subsequently, under the guidance of classification scores,

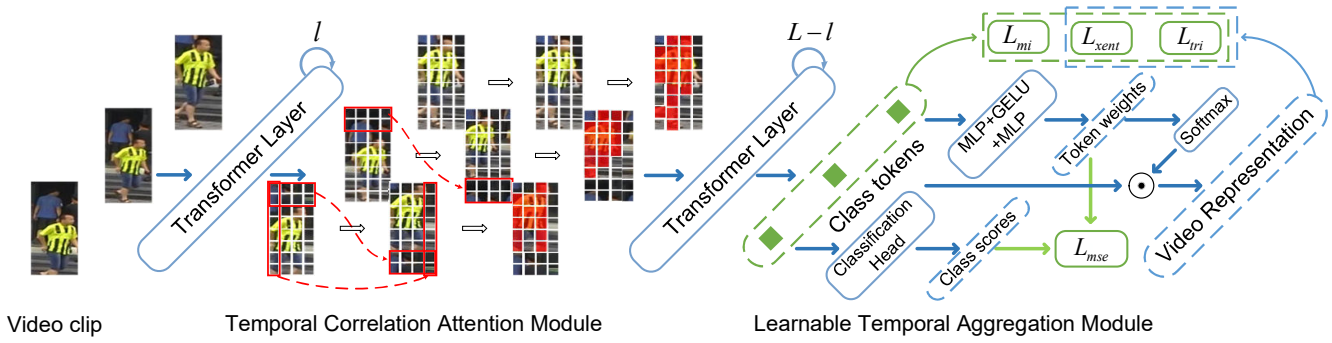

Figure 2: Framework of the proposed Temporal Correlation Vision Transformer (TCViT) for video person Re-ID.

the lightweight LTA module is introduced to weight and aggregate frame-level features.

Vision Transformer Layers

Inspired by the success of Vision Transformer (ViT) in image person Re-ID (He et al. 2021a), we use ViT to represent frames as features. Given a video sequence $\mathbb{V} = \{\boldsymbol{I}_t\}_{t=1}^{T}$ containing T frames, $\boldsymbol{I}_t \in \mathbb{R}^{H\times W\times C}$ is the t-th frame, where H, W, and C denote its height, width, and channels, respectively. The ViT encoder first splits each frame into N fixed-size patches $\boldsymbol{I}_t = \{\boldsymbol{p}_t^1, \boldsymbol{p}_t^2, \ldots, \boldsymbol{p}_t^N\}$. Then, an extra learnable class token is prepended to patches and position embedding $\mathcal{P} \in \mathbb{R}^{(N+1)\times D}$ and camera index embedding $\mathcal{C} \in \mathbb{R}^{(N+1)\times D}$ (He et al. 2021a) are applied. The input of the encoder is thus described as follows:

$$\boldsymbol{X}_t = \left[\boldsymbol{p}_t^{cls}; f\left(\boldsymbol{p}_t^1\right); f\left(\boldsymbol{p}_t^2\right); \ldots; f\left(\boldsymbol{p}_t^N\right)\right] + \mathcal{P} + \mathcal{C}, \quad (1)$$

where $f(\cdot)$ is a linear mapping function that maps each patch to a D-dimensional feature vector, and $\boldsymbol{p}_t^{cls} \in \mathbb{R}^D$ is the class token. The ViT encoder composed of L layers extracts features layer by layer. We denote the frame-level feature process by layer l as $\boldsymbol{Z}_t^l = \{\boldsymbol{z}_t^{cls}; \boldsymbol{z}_t^1; \boldsymbol{z}_t^2; \ldots; \boldsymbol{z}_t^N\}$, where $\boldsymbol{z}_t^{cls}$ and $\boldsymbol{z}_t^n$ are D-dimensional feature vectors of the class token and patches, respectively.

Temporal Correlation Attention Module

As mentioned in the Introduction, the target person is relatively static to the bounding box, providing a clue to enhance the feature of the target person. Drawing inspiration from the previous method (Hou et al. 2019), we apply temporal average pooling on frame-level features to focus on the relatively static part of the video. We then calculate the cosine similarity between the frame-level features and the pooled one to measure their temporal correlation separately. Patches with low temporal correlation correspond to relatively dynamic parts (*i.e.,* the non-target person), while those with high temporal correlation correspond to relatively static parts (*i.e.,* the target person).

However, the inadequate detector leads to misalignment between adjacent frames, resulting in an issue during temporal average pooling where the features of the target person can be confused by the misaligned background. To align the frame-level features, we use the kernelized correlation filter (KCF) (Henriques et al. 2014) to recover semantic coherence, which first calculates the cross-frame position deviation and then rolls patches in the frame-level features for alignment.

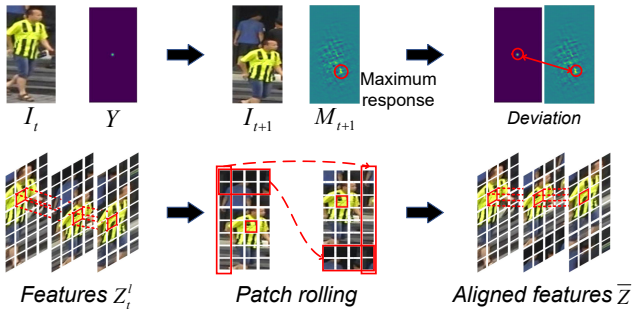

Figure 3: Diagram of Filter-based Patch Alignment

Filter-based Patch Alignment. Following KCF (Henriques et al. 2014), we transform the deviation calculation into a ridge regression problem with the kernel method. The goal is to obtain a filter with the highest response to a certain part of the target person. Therefore, the deviation across frames can be obtained by comparing the maximum response points of the filter on adjacent frames. For this reason, the first step is to calculate a filter with the highest response to the center of the frame $\boldsymbol{I}_t$.

Suppose that $\boldsymbol{Y} \in \mathbb{R}^{H\times W}$ is the response map of the filter on frame $\boldsymbol{I}_t$, as shown in the first row of Figure 3. The center of the map is the highest response, marked as one, and the response gradually decreases to zero from the center to the corners. Following KCF, we use $\boldsymbol{I}_t$ to initialize the filter and calculate the self-correlation $\boldsymbol{K}^{t,t} \in \mathbb{R}^{H\times W}$ with a Gaussian kernel as:

$$\boldsymbol{K}^{t,t} = \exp\left(-\frac{2}{\sigma^2}\left(\|\boldsymbol{I}_t\|^2 - \mathcal{F}^{-1}\left(\sum\nolimits_C(\hat{\boldsymbol{I}}_t^{\,*} \odot \hat{\boldsymbol{I}}_t)\right)\right)\right), \quad (2)$$

where $\odot$ is the element-wise product, $\mathcal{F}^{-1}(\cdot)$ denotes the inverse discrete Fourier transform, and $\sum_C(\cdot)$ means summation along the channel dimension. Moreover, σ is the bandwidth parameter, $\hat{\boldsymbol{I}}_t$ means the discrete Fourier transformation of $\boldsymbol{I}_t$, and $\hat{\boldsymbol{I}}_t^{\,*}$ is the complex conjugate of $\hat{\boldsymbol{I}}_t$.

Algorithm 1: Deviation calculation procedure

Input: Video $\mathbb{V} = \{\boldsymbol{I}_t\}_{t=1}^{T}$, response map $\boldsymbol{Y}$
Output: Deviations $(\boldsymbol{e}^x,\ \boldsymbol{e}^y)$

1: Initial filter $\boldsymbol{\alpha}$ with $\boldsymbol{I}_1$ by Eq. (3);
2: **for** $t = 2,\ t \leq T$ **do**
3: Compute the response $\boldsymbol{M}_{t+1}$ of filter $\boldsymbol{\alpha}$ on $\boldsymbol{I}_{t+1}$ by Eq. (5)
4: Obtain the deviations $(\boldsymbol{e}^x_{t+1},\ \boldsymbol{e}^y_{t+1})$ by Eq. (6)
5: Based on deviations, roll the patches of feature $\boldsymbol{Z}^l_t$ for alignment.
6: **end for**
7: **return** Aligned features $\bar{\boldsymbol{Z}}^l$

Based on $\boldsymbol{K}^{t,t}$ and the regression target $\boldsymbol{Y}$, the filter $\boldsymbol{\alpha}$ can be computed as follows:

$$\boldsymbol{\alpha} = \frac{\mathcal{F}(\boldsymbol{Y})}{\mathcal{F}(\boldsymbol{K}^{t,t}) + \lambda}, \tag{3}$$

where $\mathcal{F}(\cdot)$ represents the discrete Fourier transformation and λ is a regularization parameter. The filter $\boldsymbol{\alpha} \in \mathbb{R}^{H \times W}$ is most responsive to the center of $\boldsymbol{I}_t$. To obtain the deviation across frames, we compute the response map of the filter on the next frame $\boldsymbol{I}_{t+1}$. We first compute the cross-correlation $\boldsymbol{K}^{t,t+1}$ between $\boldsymbol{I}_t$ and $\boldsymbol{I}_{t+1}$ as follows:

$$\boldsymbol{K}^{t,t+1} = exp\Bigg(-\frac{1}{\sigma^2}\Big(\|\boldsymbol{I}_t\|^2 + \|\boldsymbol{I}_{t+1}\|^2 - 2\mathcal{F}^{-1}\Big(\sum\nolimits_C (\hat{\boldsymbol{I}}_t^{\,*} \odot \hat{\boldsymbol{I}}_{t+1}) \Big) \Big) \Bigg). \tag{4}$$

Then, the response $\boldsymbol{M}_{t+1} \in \mathbb{R}^{H \times W}$ of the filter $\boldsymbol{\alpha}$ on frame $\boldsymbol{I}_{t+1}$ can be calculated as follows:

$$\boldsymbol{M}_{t+1} = \mathcal{F}^{-1}\left(\mathcal{F}(\boldsymbol{K}^{t,t+1}) \odot \boldsymbol{\alpha}\right). \tag{5}$$

The location of the maximum response point of $\boldsymbol{M}_{t+1}$, denoted as $(\boldsymbol{M}^x_{t+1}, \boldsymbol{M}^y_{t+1})$, corresponds to the center of frame $\boldsymbol{I}_t$. Accordingly, the deviation can be calculated as:

$$(\boldsymbol{e}^x_{t+1},\ \boldsymbol{e}^y_{t+1}) = (\boldsymbol{M}^x_{t+1} - W/2,\ \boldsymbol{M}^y_{t+1} - H/2), \tag{6}$$

where $\boldsymbol{e}^x_{t+1}$ and $\boldsymbol{e}^y_{t+1}$ are horizontal and vertical deviations, respectively. Based on it, TCA rolls patches (except for class token) in feature maps for alignment, as shown in the second row of Figure 3. This operation ensures that frame-level feature maps have the same semantics along the same spatial region, thus restoring semantic coherence. The entire process of the deviation calculation is outlined in Algorithm 1 to facilitate comprehension.

Correlation Attention. As illustrated in Figure 4, the features aligned after layer l are represented as $\bar{\boldsymbol{Z}}^l_t$, and temporal average pooling is applied to them to obtain $\boldsymbol{V} \in \mathbb{R}^{N \times D}$:

$$\boldsymbol{V} = \frac{1}{T}\sum_{t=1}^{T} \bar{\boldsymbol{Z}}^l_t. \tag{7}$$

As mentioned above, averaging over the temporal dimension focuses on the relatively static parts. We can determine whether a patch is important (*i.e.*, belonging to the target person) by comparing it with the pooled ones. To do this, we calculate the cosine similarity between $\bar{\boldsymbol{Z}}^l_t$ and $\boldsymbol{V}$ as follows:

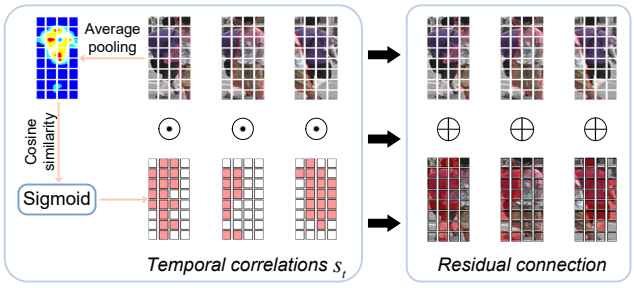

Figure 4: Diagram of Correlation Attention

$$\boldsymbol{s}_t = \frac{\bar{\boldsymbol{Z}}^l_t}{\|\bar{\boldsymbol{Z}}^l_t\|} \otimes \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|}, \tag{8}$$

where $\otimes$ denotes the inner product. Here, $\boldsymbol{s}_t \in \mathbb{R}^N$, which lies between 0 and 1, estimates the temporal correlation of each patch. Applying $\mathrm{Sigmoid}(\boldsymbol{s}_t)$ to the input features, TCA will return frame-level features that focus more on the target person as follows:

$$\boldsymbol{Z}^l_t = \bar{\boldsymbol{Z}}^l_t + \bar{\boldsymbol{Z}}^l_t \odot \mathrm{Sigmoid}(\boldsymbol{s}_t). \tag{9}$$

During the alignment operation, we initial filter with the first frame. However, in some cases, the first frame may not be an accurate detection result. Aligning subsequent frames to it can cause the body structure of the target person to be distorted (*e.g.*, patches of the legs rolling up to the upper body). Our aim is to make sure that the features have the same semantic in the same spatial region so that patches with the same semantic can be merged together during temporal average pooling. Therefore, the disruption of body structure does not reduce performance, and the original patch arrangement will be restored for the subsequent layer after the TCA process.

Learnable Temporal Aggregation Module

ViT, when used in conjunction with the TCA module, encodes each frame as a frame-level feature. In the traditional approach, the class tokens $\boldsymbol{z}^{\text{cls}}_t$ are separated from $\boldsymbol{Z}^L_t$ and then averaged over time to generate the final video representation. However, in crowded scenes, some frames may be occluded, resulting in only a partial view of the target person. It is more appropriate to assign a higher weight to complete frames compared to occluded ones, rather than treating all frames equally. To this end, a lightweight LTA module is introduced to weigh and aggregate each frame-level feature.

The illustration in Figure 2 provides an overview of our LTA. Specifically, the class tokens $\boldsymbol{Z}^{\text{cls}} \in \mathbb{R}^{T \times D}$ are separated from the output of the last layer in ViT. Subsequently, a linear layer and a RELU layer are used to downsample the channel dimension:

$$\boldsymbol{F} = \mathrm{GELU}(\mathrm{MLP}(\boldsymbol{Z}^{\text{cls}})). \tag{10}$$

Here, $\boldsymbol{F} \in \mathbb{R}^{T\times D/r}$ compresses the information of class tokens by downsampling it at a rate of r. We reshape $\boldsymbol{F}$ into a vector $\boldsymbol{f} \in \mathbb{R}^{T\cdot D/r}$ and compute the weight $\boldsymbol{w} \in \mathbb{R}^{T}$ of class tokens as follows:

$$\begin{aligned} \boldsymbol{w} &= \mathrm{MLP}(\boldsymbol{f}), \\ \boldsymbol{v} &= \frac{1}{T}\sum\nolimits_{T}(\boldsymbol{Z}^{\mathrm{cls}} \odot \mathrm{Sigmoid}(\boldsymbol{w})), \end{aligned} \tag{11}$$

where $\sum_T(\cdot)$ means summation along time dimension. Subsequently, LTA multiplies $\boldsymbol{Z}^{cls}$ with $\mathrm{Sigmoid}(\boldsymbol{w})$ and aggregates over time to obtain the final video representation $\boldsymbol{v} \in \mathbb{R}^{D}$.

It is insufficient to learn the weights only through cascaded linear layers. To better guide the training of LTA, we calculate classification scores for $\boldsymbol{Z}^{\mathrm{cls}}$ with a classification head and use the cross entropy loss as the classification loss for training. Generally, partially occluded frames have lower classification scores due to incomplete information about the target person, while complete frames tend to have higher classification scores. Taking this into account, we apply a Mean Squared Error (MSE) loss to measure the discrepancy between $\boldsymbol{x}$ and the classification score:

$$\begin{aligned} \boldsymbol{c} &= \mathrm{MLP}(\boldsymbol{Z}^{\mathrm{cls}}), \\ \mathcal{L}_{mse}(\boldsymbol{c}, \boldsymbol{w}) &= \sum\nolimits_{T} \|y(\boldsymbol{c}) - \boldsymbol{w}\|_2, \end{aligned} \tag{12}$$

where $\boldsymbol{c} \in \mathbb{R}^{T\times M}$ is the classification score, M is the number of classes. The function $y(\boldsymbol{c}) \in \mathbb{R}^{T}$ selects a score corresponding to the ground truth from each row of $\boldsymbol{c}$.

Last, with the help of the classification scores, LTA can efficiently aggregate frame-level features to create an informative representation of the video.

Objective Function

We adopt the triplet loss (Hermans, Beyer, and Leibe 2017) $\mathcal{L}_{tri}$ and mutual information loss (Hjelm et al. 2019) $\mathcal{L}_{mi}$ to guide the training of frame-level feature maps, similar to (Bai et al. 2022; Yan et al. 2020). Additionally, we apply the MSE loss $\mathcal{L}_{mse}$ in Eq. (12) and cross entropy loss $\mathcal{L}_{xent}$ to improve the learned weights:

$$\mathcal{L}_f = \mathcal{L}_{tri}(\boldsymbol{Z}^{\mathrm{cls}}) + \mathcal{L}_{mi}(\boldsymbol{Z}^{\mathrm{cls}}) + \mathcal{L}_{xent}(\boldsymbol{c}) + \mathcal{L}_{mse}(\boldsymbol{c}, \boldsymbol{w}). \tag{13}$$

After the process of LTA, $\mathcal{L}_{tri}$ and $\mathcal{L}_{xent}$ are adopted to guide the training of the video representation $\boldsymbol{v}$:

$$\mathcal{L}_v = \mathcal{L}_{tri}(\boldsymbol{v}) + \mathcal{L}_{xent}(\boldsymbol{v}). \tag{14}$$

The overall objective function of our TCF is a combination of the above two losses:

$$\mathcal{L} = \mathcal{L}_v + \frac{\|\mathcal{L}_v\|}{\|\mathcal{L}_f\|}\mathcal{L}_f. \tag{15}$$

Experiments

Datasets and Settings

Datasets. We evaluate our method on three datasets, *i.e.*, MARS (Zheng et al. 2016), LS-VID (Li et al. 2019), and iLiDS-VID (Wang et al. 2014), for video person Re-ID and one dataset, *i.e.*, VVeRI-901-trial (Zhao et al. 2021) for video vehicle Re-ID. More details are introduced in supplementary materials.

Implementation Details. The ViT-base ($l = 12$) is adopted as the baseline for our TCViT, the same as the CAViT (Wu et al. 2022). During training, we randomly select 8 identities and sample 4 sequences for each identity. We follow the restricted random sampling strategy (Li et al. 2018), which evenly divides the video sequence into 8 chunks and randomly selects one frame per chunk. The frames are resized to 256×128 and augmented by random flipping and erasing (Wang et al. 2018; Zhong et al. 2020). We set the reduction rate $r = 8$. Additionally, we add two TCA modules after the 7th and the 9th ViT layers separately. The Adam (Kingma and Ba 2014) optimizer with a weight decay of 0.001 is used as optimizer, and the learning rate is initialized as 0.0005 with a cosine learning rate scheduler. The model is trained for 90 epochs. During evaluation, we split each video sequence into multiple 8-frame video clips and obtain the video-level representation by averaging all extracted clip features. The mean Average Precision (mAP) and the Cumulative Matching Characteristics (CMC) are adopted as evaluation metrics.

Methods	GFs	Para.	Infer. Time	LS-VID mAP	LS-VID rank-1
base.	11.04	86.3	1.0x	80.5	88.2
+TCA w/o A.	11.04	86.3	1.03x	81.0	88.7
+TCA	11.04	86.3	1.40x	81.4	89.2
+LTA w/o M.	11.05	87.2	1.02x	81.0	88.8
+LTA	11.05	87.2	1.02x	82.0	89.3
TCViT	11.05	87.2	1.44x	**83.1**	**90.1**

Table 1: Component analysis of TCViT on LS-VID. A. and M. means align operation and MSE loss separately.

Ablation Study

For a fair comparison, we build the baseline by degrading TCViT without TCA and LTA, which we refer to as "base." for simplicity. In addition, we compare the influence of TCA and LTA on inference time in Table 1.

Component Analysis of TCA. We evaluate the contributions of TCA on LS-VID in Table 1. TCA is parameter-free and does not add any computational costs, resulting in a 0.9% / 1.0% increase in rank-1/ mAP over the baseline. In particular, filter-based patch alignment is essential for TCA, and its absence leads to a decrease in performance.

We conduct experiments to investigate the organizational rationality of TCA by inserting two TCA modules into different positions. "$\boldsymbol{Z}^3+\boldsymbol{Z}^5$" means respectively inserting two TCA modules after the 3rd and 5th ViT layers. The combination of "$\boldsymbol{Z}^7 + \boldsymbol{Z}^9$" achieves the best performance, as shown in Table 2 (a). In shallow layers, patches learn features from similar patches surrounding them. TCA weakens non-target persons' and background patches, which hinders the relation

Methods	mAP	rank-1
TCViT	83.1	90.1
$Z^3 + Z^5$	76.6	84.5
$Z^5 + Z^7$	81.7	88.6
$Z^7 + Z^9$	**83.1**	**90.1**
$Z^9 + Z^{11}$	82.7	89.6

(a) The results of TCViT with two TCA modules at different positions.

Methods	mAP	rank-1
TCViT	83.1	90.1
$Z^7 + Z^9 + Z^{11}$	83.0	89.9
$Z^7 + Z^9$	**83.1**	**90.1**
$Z^7 + Z^8$	82.8	89.8
Z^7	82.6	89.3

(b) The results of TCViT with a different number of TCA modules.

Methods	Para.	mAP	rank-1
TCViT	87.02	83.1	90.1
$r = 2$	87.24	82.1	89.1
$r = 4$	87.09	82.6	89.6
$r = 8$	**87.02**	**83.1**	**90.1**
$r = 16$	86.98	82.9	89.8

(c) The results of TCViT with different reduction rate r in the LTA module.

Table 2: Ablation study on LS-VID. More details are explained in the text.

Methods	Proc.	LS-VID		MARS		iLiDS-VID	
		mAP	rank-1	mAP	rank-1	rank-1	rank-5
GLTR(Li et al. 2019)	ICCV2019	44.3	63.1	78.5	87.0	86.0	98
VRSTC(Hou et al. 2019)	CVPR2019	-	-	82.3	88.5	83.4	95.5
M3D(Li, Zhang, and Huang 2019)	AAAI2019	40.1	57.7	74.1	84.4	74.0	94.3
AP3D(Gu et al. 2020)	ECCV2020	73.2	84.5	85.1	90.1	88.7	-
TCLNet(Hou et al. 2020)	ECCV2020	70.3	81.5	85.1	89.8	86.6	-
AFA(Chen et al. 2020)	ECCV2020	-	-	82.9	90.2	88.5	96.8
MGH(Yan et al. 2020)	CVPR2020	-	-	85.8	90.0	85.6	97.1
MG-RAFA(Zhang et al. 2020)	CVPR2020	-	-	85.9	88.8	88.6	98
BiCnet-TKS(Hou et al. 2021)	CVPR2021	75.1	84.6	86.0	90.2	-	-
GRL(Liu et al. 2021b)	CVPR2021	-	-	84.8	91.0	96.7	98.3
STRF(Aich et al. 2021)	ICCV2021	-	-	86.1	90.3	89.3	-
STMN(Eom et al. 2021)	ICCV2021	69.2	82.1	84.5	90.5	-	-
DenseIL(He et al. 2021b)	ICCV2021	-	-	87.0	90.8	92.0	98.0
SINet(Bai et al. 2022)	CVPR2022	79.6	87.4	86.2	91.0	92.5	-
CAViT(Wu et al. 2022)	ECCV2022	79.2	89.2	87.2	90.8	93.3	98.0
TCViT	-	**83.1**	**90.1**	**87.6**	**91.7**	**94.3**	**99.3**

Table 3: Performance comparison with state-of-the-art methods on LS-VID, MARS, and iLiDS-VID.

Methods	Sequence Length	Proc.	VVeRI-901	
			mAP	rank-1
base.	8	ICLR2021	63.5	56.7
BiCnet-TKS	8	CVPR 2021	50.8	41.3
AP3D	4	ECCV 2020	61.2	52.5
Token shift	8	ICCV 2019	67.4	57.5
CAViT	8	ECCV 2022	65.6	60.0
TCViT	4	-	65.0	58.7
	8		**69.0**	**63.9**

Table 4: Comparison with state-of-the-arts on VVeRI-901.

construction between patches in shallow layers. However, in deep layers, only a few patches with important semantic information are taken into account. Therefore, TCA improves performance by weighting the deep frame-level features to focus on the target person, which is in line with the findings of Chang et al. (Chang et al. 2023).

We also evaluate the influence of the number of TCA modules, as illustrated in Table 2 (b). It can be seen that the performance increases and reaches its peak with two separate TCA modules after the 7th and 9th layers. Stacking three TCA modules does not bring any performance gain, and "$Z^7 + Z^9$" provides the best balance between performance and complexity.

Component Analysis of LTA. We evaluate the effectiveness of LTA, and the results are presented in Table 1. Replacing the temporal average pooling with LTA has increased the rank-1/mAP by 0.5%/0.6% over the baseline. Additionally, when MSE loss is embedded to guide LTA training, performance is further improved by 1.0%/0.5% rank-1/mAP on LS-VID. LTA is a lightweight module that only introduces 1% extra parameters and negligible GFLOPs, yet has achieved a total improvement of 1.5%/1.1% rank-1/mAP.

As shown in Eq. (10), LTA compresses the class tokens at a rate of r, and Table 2 (c) presents the results on LS-VID at different reduction rates r, *i.e.*, $r = 2/4/8/16$. As the reduction rate increases, the performance improves until it reaches its peak at $r = 8$. After that, further increasing the rate does not bring any benefit, probably due to the information loss caused by the excessive reduction rate.

(a) Visualization of Filter-based Patch Alignment. The first row is an example of misaligned videos, while the second is the video sequence after alignment.

(b) Visualization of Correlation Attention. The first row is two examples of videos under occlusion, while the second is the corresponding temporal correlation $\boldsymbol{s}_t$.

Figure 5: Visualization of Filter-based Alignment and Correlation Attention in TCA.

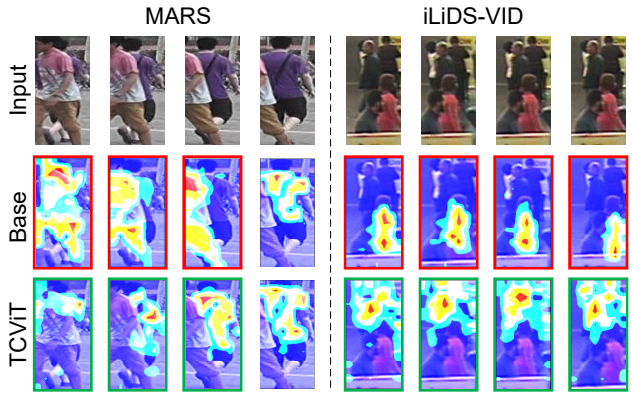

Figure 6: Visualization of the difference between the baseline and our TCViT framework.

Comparison with State-of-the-art Methods

In Table 3, we compare TCViT with existing state-of-the-art methods on three prevalent video person Re-ID benchmarks. **(1)** LS-VID is one of the most challenging datasets. Our approach achieves 83.1% in mAP, surpassing the previous best method CAViT (Wu et al. 2022) by a considerable margin. **(2)** On MARS, we obtain a rank-1/mAP of 91.7%/87.6%, outperforming all other state-of-the-art methods. Most of the methods are CNN-based, except CAViT (Wu et al. 2022) and DenseIL (He et al. 2021b). Our method increases rank-1 by 0.9% over the above two methods. As Wu et al. (Wu et al. 2022) pointed out, the ID switch caused by the GMMCP tractor (Dehghan, Modiri Assari, and Shah 2015) and the misalignment caused by the DPM detector (Felzenszwalb et al. 2009) led to confusion in the video Re-ID task, resulting in a performance bottleneck on MARS. Nevertheless, TCViT achieves the best performance among all other methods. **(3)** iLiDS-VID is a small benchmark with only 500 identities. TCViT still performs well on the small-scale dataset, showing robustness and effectiveness.

To test the generalizability of our TCViT, we conduct experiments on VVeRI-901-trial (Zhao et al. 2021). As Table 4 shows, TCViT outperforms the previous best methods AP3D and CAViT by 6.2%/3.8% rank-1/mAP and 3.9%/3.4% rank-1/mAP, respectively, when using the same sequence length. Additionally, the combination of TCA and LTA results in a 7.3%/5.5% rank-1/mAP improvement over the baseline.

Visualization Analysis

Visualization of Filter-based Patch Alignment. We visualize the results of the alignment operation. As shown in Figure 5 (a), the misaligned frames in the input sequence are highlighted with red boxes. After the alignment operation, the misaligned frames align with each other, as indicated by the green boxes. This demonstrates that our TCA model accurately calculates the deviation of the target person between frames and restores temporal coherence.

Visualization of Temporal Correlations. In Figure 5 (b), we visualize the temporal correlations of frames with occlusion. We sample two identities from the MARS dataset. The top row shows the raw frames with the non-target person occlusion, while the bottom row shows the corresponding temporal correlation $\boldsymbol{s}_t$ in Eq. (8). It is clear that when the non-target persons are relatively dynamic to the bounding box, patches belonging to them are assigned lower weights. In contrast, the attention map is activated when the relatively static part (*e.g.*, target person) appears. This demonstrates that correlation attention is prone to focus on the relatively static part. When the non-target person comes from another direction or passes the target person, TCA can weaken the occluded patches and enhance the valuable ones.

Visualization of Activation Maps. We further evaluate the performance of our method by comparing the activation map with the baseline on MARS and iLiDS-VID. As marked by red boxes in the second row of Figure 6, the baseline model cannot distinguish the target person from the non-target person, while our TCViT framework can focus on the target person even under severe occlusion.

Conclusion

This paper aims to develop a better representation of the target person, particularly when they are partially occluded from view. Our TCViT utilizes the relative state to differentiate the target person from occlusion and aggregates frame-level features based on completeness. With two lightweight TCA and LTA modules, our TCViT achieves competitive performance.

Acknowledgements

This work was supported partly by the National Key R&D Program of China under Grant 2021YFB1714700, NSFC under Grants 62088102 and 62106192, Natural Science Foundation of Shaanxi Province under Grant 2022JC41, and Fundamental Research Funds for the Central Universities under Grant XTR042021005.

References

Aich, A.; Zheng, M.; Karanam, S.; Chen, T.; Roy-Chowdhury, A. K.; and Wu, Z. 2021. Spatio-temporal representation factorization for video-based person re-identification. In *ICCV*, 152–162.

Bai, S.; Ma, B.; Chang, H.; Huang, R.; and Chen, X. 2022. Salient-to-Broad Transition for Video Person Re-Identification. In *CVPR*, 7339–7348.

Bolme, D. S.; Beveridge, J. R.; Draper, B. A.; and Lui, Y. M. 2010. Visual object tracking using adaptive correlation filters. In *CVPR*, 2544–2550.

Chang, S.; Wang, P.; Lin, M.; Wang, F.; Zhang, D. J.; Jin, R.; and Shou, M. Z. 2023. Making Vision Transformers Efficient from A Token Sparsification View. In *CVPR*, 6195–6205.

Chen, D.; Li, H.; Xiao, T.; Yi, S.; and Wang, X. 2018a. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, 1169–1178.

Chen, D.; Xu, D.; Li, H.; Sebe, N.; and Wang, X. 2018b. Group consistent similarity learning via deep crf for person re-identification. In *CVPR*, 8649–8658.

Chen, G.; Rao, Y.; Lu, J.; and Zhou, J. 2020. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *ECCV*, 660–676.

Chung, D.; Tahboub, K.; and Delp, E. J. 2017. A Two Stream Siamese Convolutional Neural Network for Person Re-identification. In *ICCV*, 1983–1991.

Dai, K.; Wang, D.; Lu, H.; Sun, C.; and Li, J. 2019. Visual tracking via adaptive spatially-regularized correlation filters. In *CVPR*, 4670–4679.

Dalal, N.; and Triggs, B. 2005. Histograms of oriented gradients for human detection. In *CVPR*, 886–893.

Danelljan, M.; Shahbaz Khan, F.; Felsberg, M.; and Van de Weijer, J. 2014. Adaptive color attributes for real-time visual tracking. In *CVPR*, 1090–1097.

Dehghan, A.; Modiri Assari, S.; and Shah, M. 2015. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *CVPR*, 4091–4099.

Eom, C.; Lee, G.; Lee, J.; and Ham, B. 2021. Video-based person re-identification with spatial and temporal memory networks. In *ICCV*, 12036–12045.

Felzenszwalb, P. F.; Girshick, R. B.; McAllester, D.; and Ramanan, D. 2009. Object detection with discriminatively trained part-based models. *IEEE T-PAMI*, 32(9): 1627–1645.

Girshick, R. 2015. Fast R-CNN. In *CVPR*, 1440–1448.

Gu, X.; Chang, H.; Ma, B.; Zhang, H.; and Chen, X. 2020. Appearance-preserving 3D convolution for video-based person re-identification. In *ECCV*, 228–243.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021a. TransReID: Transformer-based object re-identification. In *ICCV*, 15013–15022.

He, T.; Jin, X.; Shen, X.; Huang, J.; Chen, Z.; and Hua, X.-S. 2021b. Dense Interaction Learning for Video-based Person Re-identification. In *ICCV*, 1490–1501.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2012. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, 702–715.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2014. High-speed tracking with kernelized correlation filters. *IEEE T-PAMI*, 37(3): 583–596.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Hou, R.; Chang, H.; Ma, B.; Huang, R.; and Shan, S. 2021. BiCnet-TKS: Learning efficient spatial-temporal representation for video person re-identification. In *CVPR*, 2014–2023.

Hou, R.; Chang, H.; Ma, B.; Shan, S.; and Chen, X. 2020. Temporal complementary learning for video person re-identification. In *ECCV*, 388–405.

Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; and Chen, X. 2019. VRSTC: Occlusion-free video person re-identification. In *CVPR*, 7183–7192.

Kalayeh, M. M.; Basaran, E.; Gökmen, M.; Kamasak, M. E.; and Shah, M. 2018. Human semantic parsing for person re-identification. In *CVPR*, 1062–1071.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Li, J.; Wang, J.; Tian, Q.; Gao, W.; and Zhang, S. 2019. Global-local temporal representations for video person re-identification. In *ICCV*, 3958–3967.

Li, J.; Zhang, S.; and Huang, T. 2019. Multi-scale 3D convolution network for video based person re-identification. In *AAAI*, 8618–8625.

Li, S.; Bak, S.; Carr, P. W.; and Wang, X. 2018. Diversity Regularized Spatiotemporal Attention for Video-based Person Re-identification. In *CVPR*, 369–378.

Liu, J.; Zha, Z.-J.; Wu, W.; Zheng, K.; and Sun, Q. 2021a. Spatial-Temporal Correlation and Topology Learning for Person Re-Identification in Videos. In *CVPR*, 4370–4379.

Liu, X.; Zhang, P.; Yu, C.; Lu, H.; and Yang, X. 2021b. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *CVPR*, 13334–13343.

Ma, C.; Huang, J.-B.; Yang, X.; and Yang, M.-H. 2015. Hierarchical convolutional features for visual tracking. In *CVPR*, 3074–3082.

McLaughlin, N.; del Rincon, J. M.; and Miller, P. 2016. Recurrent Convolutional Network for Video-Based Person Re-identification. In *CVPR*, 1325–1334.

Song, C.; Huang, Y.; Ouyang, W.; and Wang, L. 2018. Mask-guided contrastive attention model for person re-identification. In *CVPR*, 1179–1188.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *ECCV*, 480–496.

Wang, T.; Gong, S.; Zhu, X.; and Wang, S. 2014. Person re-identification by video ranking. In *ECCV*, 688–703.

Wang, Y.; Wang, L.; You, Y.; Zou, X.; Chen, V.; Li, S.; Huang, G.; Hariharan, B.; and Weinberger, K. Q. 2018. Resource aware person re-identification across multiple resolutions. In *CVPR*, 8042–8051.

Wu, J.; He, L.; Liu, W.; Yang, Y.; Lei, Z.; Mei, T.; and Li, S. Z. 2022. CAViT: Contextual Alignment Vision Transformer for Video Object Re-identification. In *ECCV*, 549–566.

Xu, Z.; Hrustic, E.; and Vivet, D. 2020. Centernet heatmap propagation for real-time video object detection. In *ECCV*, 220–234.

Yan, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; Tai, Y.; and Shao, L. 2020. Learning Multi-Granular Hypergraphs for Video-Based Person Re-Identification. In *CVPR*, 2899–2908.

Yang, J.; Zheng, W.-S.; Yang, Q.; Chen, Y.-C.; and Tian, Q. 2020. Spatial-Temporal Graph Convolutional Network for Video-Based Person Re-Identification. In *CVPR*, 3289–3299.

Zhang, K.; Zhang, L.; Liu, Q.; Zhang, D.; and Yang, M.-H. 2014. Fast visual tracking via dense spatio-temporal context learning. In *ECCV*, 127–141.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2019. Densely semantically aligned person re-identification. In *CVPR*, 667–676.

Zhang, Z.; Lan, C.; Zeng, W.; and Chen, Z. 2020. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *CVPR*, 10407–10416.

Zhao, J.; Qi, F.; Ren, G.; and Xu, L. 2021. PhD Learning: Learning with Pompeiu-hausdorff Distances for Video-based Vehicle Re-Identification. In *CVPR*, 2225–2235.

Zheng, L.; Bie, Z.; Sun, Y.; Wang, J.; Su, C.; Wang, S.; and Tian, Q. 2016. MARS: A video benchmark for large-scale person re-identification. In *ECCV*, 868–884.

Zheng, M.; Karanam, S.; Wu, Z.; and Radke, R. J. 2019. Re-identification with consistent attentive siamese networks. In *CVPR*, 5735–5744.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *AAAI*, 13001–13008.

Zhou, Z.; Huang, Y.; Wang, W.; Wang, L.; and Tan, T. 2017. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 4747–4756.