Multi-Stage Spatio-Temporal Aggregation Transformer for Video Person Re-identification

Ziyi Tang, Ruimao Zhang, *Member, IEEE,* Zhanglin Peng, Jinrui Chen, Liang Lin, *Senior Member, IEEE*

***Abstract*—In recent years, the Transformer architecture has shown its superiority in the video-based person re-identification task. Inspired by video representation learning, these methods mainly focus on designing modules to extract informative spatial and temporal features. However, they are still limited in extracting local attributes and global identity information, which are critical for the person re-identification task. In this paper, we propose a novel Multi-Stage Spatial-Temporal Aggregation Transformer (MSTAT) with two novel designed proxy embedding modules to address the above issue. Specifically, MSTAT consists of three stages to encode the attribute-associated, the identity-associated, and the attribute-identity-associated information from the video clips, respectively, achieving the holistic perception of the input person. We combine the outputs of all the stages for the final identification. In practice, to save the computational cost, the Spatial-Temporal Aggregation (STA) modules are first adopted in each stage to conduct the self-attention operations along the spatial and temporal dimensions separately. We further introduce the Attribute-Aware and Identity-Aware Proxy embedding modules (AAP and IAP) to extract the informative and discriminative feature representations at different stages. All of them are realized by employing newly designed self-attention operations with specific meanings. Moreover, temporal patch shuffling is also introduced to further improve the robustness of the model. Extensive experimental results demonstrate the effectiveness of the proposed modules in extracting the informative and discriminative information from the videos, and illustrate the MSTAT can achieve state-of-the-art accuracies on various standard benchmarks.**

***Index Terms*—Video-based Person Re-ID, Transformer, Spatial Temporal Modeling, Deep Representation Learning**

I. Introduction

PERSON Re-identification (re-ID) [4], [24], [26], which aims at matching pedestrians across different camera views at different times, is a critical task of visual surveillance. In the earlier stage, the studies have mainly focused on image-based person re-ID [24], [26], [45], which mine the discriminative information in the spatial domain. With the development of the monitoring sensors, multi-modality information has been introduced to re-ID task [31], [70], [71]. Numerous methods have been proposed to break down barriers between modalities regarding their image styles [86], structural features [81], [84], [97], or network parameters [31], [82].

On the other hand, some studies have exploited multi-frame data and proposed various schemes [38], [61], [100] to extract informative temporal representations to pursue video-based person re-ID. In such a setting, each time a non-labeled query tracklet clip is given, its discriminative feature representation needs to be extracted to retrieve the clips of the corresponding person in the non-labeled gallery. In practice, how to simultaneously extract such discriminative information

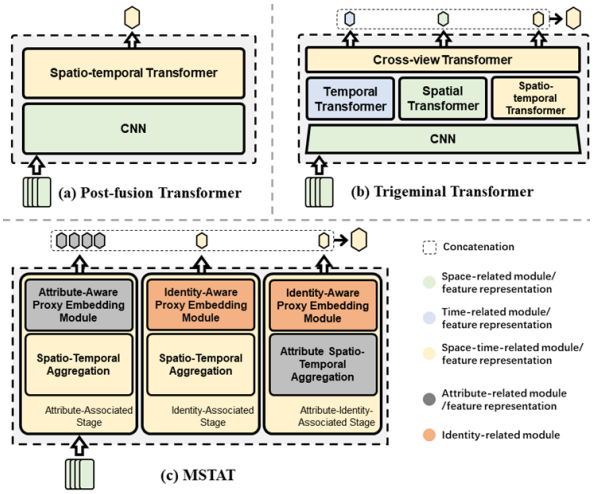


Fig. 1: Comparison between different Transformer-based frameworks for video re-ID. (a) shows the framework where the Transformer fuse post-CNN features of the entire video. (b) is Trigeminal Transformer [50], including three separate streams for temporal, spatial, and spatio-temporal feature extraction. (c) displays a multi-stage spatio-temporal aggregation Transformer, which consists of three stages, all with a spatio-temporal view but different meanings.

Ziyi Tang, Ruimao Zhang, and Jinrui Chen are with The Chinese University of Hong Kong (Shenzhen), and Ziyi Tang is also with Sun Yat-sen University (e-mail: tangziyi@cuhk.edu.cn, ruimao.zhang@ieee.org, and 120090765@link.cuhk.edu.cn).

Zhanglin Peng is with the Department of Computer Science, The University of Hong Kong, Hong Kong, China (e-mail: zhanglin.peng@connect.hku.hk).

Liang Lin is with the School of Computer Science and Engineering, Sun Yat-sen University (e-mail: linliang@ieee.org).

This paper was done when Ziyi Tang was working as a Research Assistant at The Chinese University of Hong Kong (Shenzhen).

The Corresponding Author is Ruimao Zhang

from spatial and temporal dimensions is the key to improving the accuracy of video-based re-ID.

To address such an issue, traditional methods [18] usually employ hierarchically convolutional architectures to update local patterns progressively. Furthermore, some attempts [12], [13], [47], [72], [94] adopt attention-based modules to dynamically infer discriminative information from videos. For instance, Wu *et al.* [71] embed body part prior knowledge inside the network architecture via dense and non-local region-based attention. Although recent years have witnessed the success of convolution-based methods [10], [11], [18], [36], [42], [73], [94], [104], they have encountered a bottleneck of accuracy improvement, as convolution layers suffer from their intrinsic limitations of spatial-temporal dependency modeling and information aggregation [96].

Recently, the Transformer architecture [22], [30], [53], [89] has attracted much attention in the computer vision area because of its excellent context modeling ability. The core idea of such a model is to construct interrelationships between local contents via global attention operation. In the literature, some hybrid network architectures [17], [32], [50] have been proposed to tackle long-range context modeling in video-based re-ID. A widely used paradigm is to leverage Transformer as the post-processing unit, coupled with a convolutional neural network (CNN) as the basic feature extractor. For example, as summarized in Fig. 1 (a), He *et al.* [33] and Zhang *et al.* [95] adopt a monolithic Transformer to fuse frame-level CNN feature. As shown in Fig. 1 (b), Liu *et al.* [50] take a step further and put forward multi-stream Transformer architecture in which each stream emphasizes a particular dimension of the video features. In a hybrid architecture, however, the 2D CNN bottom encoder restricts the long-range spatio-temporal interactions among local contents, which hinders the discovery of contextual cues. Later, to address this problem, some pure Transformer-based approaches are introduced to video-based re-ID. Nevertheless, the existing Transformer-based frameworks are mainly motivated by those in video understanding and concentrate on designing the architecture to learn spatial-temporal representations efficiently. Most of them are still limited in extracting informative and human-relevant discriminative information from the video clips, which are critical for large-scale matching tasks [37], [92], [98], [104].

To address the above issues, we propose a novel Multi-stage Spatial-Temporal Aggregation Transformer framework, named MSTAT, which consists of **three stages** to respectively encode the attribute-associated, the identity-associated, and the attribute-identity-associated information from video clips. Firstly, to save the computational cost, the **Spatial-Temporal Aggregation** (STA) modules [2], [5] are firstly adopted in each stage as their building blocks to conduct the self-attention operations along the spatial and temporal dimensions separately. Further, as shown in Fig. 1, we introduce the plug-and-play **Attribute-Aware Proxy** and **Identity-Aware Proxy** (AAP and IAP) embedding modules into different stages, for the purpose of reserving informative attribute features and aggregating discriminative identity features respectively. They are both implemented by self-attention operations but with different learnable proxy embedding schemes. For the AAP embedding module, AAPs play the role of attribute queries to reserve a diversity of implicit attributes of a person. Arguably, the combination of these attribute representations is informative and provides discriminative power, complementary to the identity-only prediction. In contrast, the IAP embedding module maintains a group of IAPs as key-value pairs. With explicit constraints, they learn to successively match and aggregate the discriminative identity-aware features embedded in patch tokens. During similarity measurement, the output feature representations of the three stages are concatenated to form a holistic view of the input person.

In practice, a Transformer-specific data augmentation scheme, Temporal Patch Shuffling, is also introduced, which randomly rearranges the patches temporally. With such a scheme, the enriched training data effectively improve the ability to learn invariant appearance features, leading to the robustness of the model. Extensive experiments on three public benchmarks demonstrate our proposed framework is superior to the state-of-the-art on different metrics. Concretely, we achieve the best performance of 91.8% rank-1 accuracy on MARS, which is the largest video re-ID dataset at present.

In summary, our contributions are three-fold. (1) We introduce a Multi-stage Spatial-Temporal Aggregation Transformer framework (MSTAT) for video-based person re-ID. Compared to existing Transformer-based frameworks, MSTAT better learns informative attribute features and discriminative identity features. (2) For different stages, we devise two different proxy embedding modules, named Attribute-Aware and Identity-Aware Proxy embedding modules, to extract informative attribute features and aggregate discriminative identity features from the entire video, respectively. (3) A simple yet effective data augmentation scheme, referred to as Temporal Patch Shuffling, is proposed to consolidate the network's invariance to appearance shifts and enrich training data.

II. Related Works

A. Image-Based Person Re-ID

Image-based person re-ID mainly focuses on person representation learning. Early works focus primarily on carefully designed handcraft features [4], [24], [26], [43], [45], [103]. Recently, The flourishing deep learning has become the mainstream method for learning

representation in person ReID [42], [64], [66], [73], [76], [88]. For the last few years, CNN has been a widely-used feature extractor [15], [39], [42]–[44], [64], [75], [80], [94]. OSNet [104] fuses multi-scale features in an attention-style sub-network to obtain informative omni-scale features. Some works [16], [87], [98] focus on extracting and aligning semantic information to address misalignment caused by pose/viewpoint variations, imperfect person detection, etc. To avoid the misleading by noisy labels, Ye *et al.* [83] presents a self-label refining strategy, deeply integrating annotation optimization and network training. So far, some works [17], [32] also explore Image-based person re-ID based on Vision Transformer [22]. For example, TransReID [32] adopts Transformer as the backbone and extracts discriminative features from randomly sampled patch groups.

B. Video-Based Person ReID

Compared to image-based person re-ID, video-based person re-ID usually performs better because it provides temporal information and mitigates occlusion by taking advantage of multi-frame information. For capturing more robust and discriminative representation from frame sequences, traditional video-based re-ID methods usually focus on two areas: 1) encoding of temporal information; 2) aggregation of temporal information.

To encode additional temporal information, early methods [38], [61], [100] directly use temporal information as additional features. Some works [48], [54], [72], [80] use recurrent models, *e.g.*, RNNs [55] and LSTM [35], to process the temporal information. Some other works [10], [11], [52], [54], [59], [80], [105] go further by introducing the attention mechanism to apply dynamic temporal feature fusion. Another class of works [19] introduces optical flow that captures temporal motion. What is more, some works [40], [41], [62], [74], [91], [102] directly implement spatio-temporal pooling to video sequences and generate a global representation via CNNs. Recently, 3D CNNs [27], [44] learn to encode video features in a joint spatio-temporal manner. M3D [39] endows 2D CNN with multi-scale temporal feature extraction ability via multi-scale 3D convolutional kernels.

For the sake of aggregation that aims to generate discriminative features from full video features, a class of approaches [54], [93], [105] applies average pooling on the time dimension to aggregate spatio-temporal feature maps. Recently, some attention-based methods [13], [41], [71], [79] attained significant performance improvement by dynamically highlighting different video frames/regions so as to filter more discriminative features from these critical frames/regions. For instance, Liu *et al.* [50] introduce cross-attention to aggregate multi-view video features by pair-wise interaction between these views. Apart from the exploration of more effective architectural design, a branch of works study the effect of pedestrian attributes [8], [60], [101], such as *shoes, bag, and down color*, or the gait [9], [56], *i.e.* walking style of pedestrians, as a more comprehensive form of pedestrian description. Chang *et al.* [9] closely integrate two coherent tasks: gait recognition and video-based re-ID by using a hybrid framework including a set-based gait recognition branch. Some works [60], [101] embed attribute predictors into the network supported by annotations obtained from a network pretrained on an attribute dataset. Chai *et al.* [8] separate attributes into ID-relevant and ID-irrelevant ones and propose a novel pose-invariant and motion-invariant triplet loss to mine the hardest samples considering the distance of pose and motion states.

Although the above methods have made significant progress in performance, Transformer [65], which is deemed a more powerful architecture to process sequence data, may raise the performance ceiling of video-based re-ID. To illustrate this, Transformer can readily adapt to video data with the support of the global attention mechanism to capture spatio-temporal dependencies and temporal positional encoding to order spatio-temporal positions. In addition, the class token is off-the-shelf for Transformer-based models to aggregate spatio-temporal information. However, Transformer suffers from multiple drawbacks [22], [69], [89], [90], and few works have been released so far on video-based person re-ID based on Transformer. In this work, we attempt to explore the potential of intractable Transformer in video-based person re-ID.

C. Vision Transformer

Recently, Transformer has shown its ability as an alternative to CNN. Inspired by the great success of Transformer in natural language processing, recent researchers [22], [53], [53], [69] have extended Transformer to CV tasks and obtained promising results.

Bertasius $et\ al.$ [5] explores different video self-attention schemes considering their cost-performance trade-off, resulting in a conclusion that the divided space-time self-attention is optimal. Similarly, ViViT [2] factorizes self-attention to compute self-attention spatially and then temporally. Inspired by these works, we divide video self-attention into spatial attention followed by temporal attention, and we further propose a attribute-aware variant for video-based re-ID. Furthermore, little research has been done on Transformer for Video-based person re-ID. Trigeminal Transformers (TMT) [50] puts the input patch token sequence through a spatial, a temporal, and a spatio-temporal minor Transformer, respectively, and a cross-view interaction module fuses their outputs. Differently, MSTAT has three stages, all extracting spatio-temporal features but with different meanings: (1) attribute features, (2) identity features, (3) attribute-identity features.

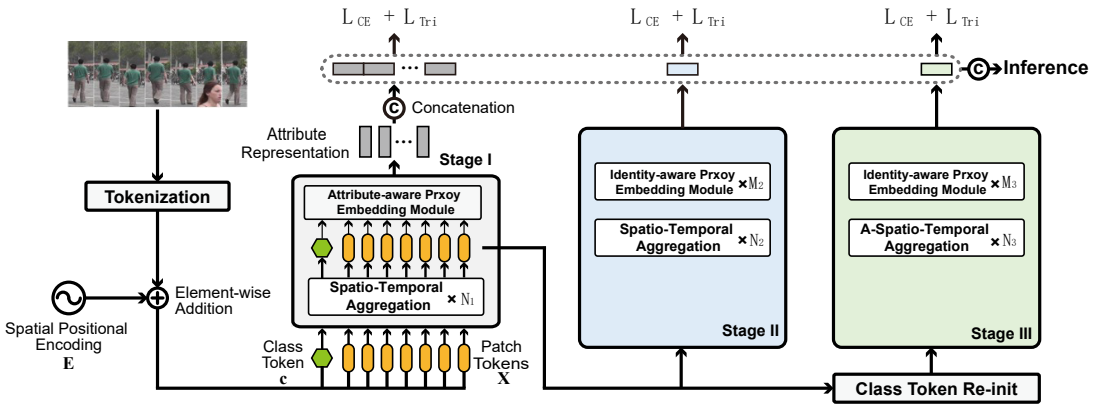


Fig. 2: The overall architecture of our proposed MSTAT which consists of three stages, all based on the Transformer architecture. **Stage I** updates the spatio-temporal patch token sequence of the input video and aggregates them into a group of attribute-associated representations. Subsequently, **Stage II** aggregates discriminative identity-associated features and **Stage III** attribute-identity-associated features, relying upon their stage-specific class tokens. Here, we omit the input and output of each module except the attribute-aware proxy embedding module in **Stage I**. At inference time, all these feature representations are combined through concatenation to infer the pedestrian's identity jointly.

III. Method

In Sec. III-A, we first overview the proposed MSTAT framework. Then, Spatio-Temporal Aggregation (STA), the normal spatial-temporal feature extractor in MSTAT, is formulated in section Sec. III-B. Along with it, we introduce the proposed Attribute-Aware Proxy (AAP) and Identity-Aware Proxy (IAP) embedding modules in Sec. III-C. Finally, Temporal Patch Shuffling (TPS), a newly introduced Transformer-specific data augmentation scheme, is presented in section III-E.

A. Overview

This section briefly summarizes the workflow of MSTAT. The overall MSTAT framework is shown in Fig. 2. Given a video tracklet $\mathbf{V} \in \mathbb{R}^{T\times 3\times H\times W}$ with T frames and the resolution of each frame is $H \times W$, the goal of MSTAT is to learn a mapping from a video tracklet $\mathbf{V}$ to a d-dimension representation space in which each identity is discriminative from the others.

Specifically, as shown on the left of Fig. 2, MSTAT first linearly projects non-overlapping image patches of size $3 \times P \times P$ into d-dimensional patch tokens, where $d = 3P^2$ denotes the embedded dimension of tokens. Thus, a patch token sequence $\mathbf{X} \in \mathbb{R}^{T\times N\times d}$ is obtained, where the number of patch tokens in each frame is denoted by $N = \frac{H\times W}{P^2}$. Meanwhile, spatial positional encoding $\mathbf{E} \in \mathbb{R}^{N\times d}$ is added to $\mathbf{X}$ in a element-wise manner for reserving spatial structure in each frame. Notably, we do not insert temporal positional encoding into $\mathbf{X}$, since the temporal order is usually not conducive to video-based re-ID, which is also demonstrated in [92]. Finally, a class token $\mathbf{c} \in \mathbb{R}^d$ is associated with $\mathbf{X}$ to aggregate global identity representation.

Next, we feed the token sequence $\mathbf{X}$ into **Stage I** of MSTAT. It takes $\mathbf{X}$ and $\mathbf{c}$ as input, and employs a stack of eight Spatio-Temporal Aggregation (STA) blocks for inter-frame and intra-frame correlation modeling. The output tokens are then fed into an Attribute-Aware Proxy (AAP) embedding module to mine rich visual attributes, a composite group of semantic cues that imply identity information, *e.g.*, `garments`, `handbags` and so on. The **Stage II** includes a series of STA blocks (three in our experiments), followed by an Identity-Aware Proxy (IAP) embedding module which is able to screen out discriminative identity-associated information by inspecting the entire sequence in parallel. In the **Stage III**, we first introduce a novel class token to directly aggregate higher-level features. In addition, a stack of Attribute-STA (A-STA) blocks is used to fuse attributes from different frames. At last, an IAP embedding module is adopted to generate a discriminative representation for the person. In the training phase, the attribute representations extracted from **Stage I** and the class tokens of **Stage II** and **Stage III** are supervised separately by a group of losses. During the testing, the attribute representations and the class tokens from the last two stages are concatenated for similarity measurement.

B. Spatio-temporal Aggregation

To begin with, we make a quick review of the vanilla Transformer self-attention mechanism first proposed in [65]. In practice, visual Transformer embeds an image into a sequence of patch tokens, and self-attention operation first linearly projects these tokens to the corresponding query $\mathbf{Q}$, key $\mathbf{K}$ and value $\mathbf{V}$ respectively. Then, the scaled product of $\mathbf{Q}$ and $\mathbf{K}$ generates an attention map $\mathbf{A}$, indicating estimated relationships between token representations in $\mathbf{Q}$ and $\mathbf{K}$. Then, $\mathbf{V}$

performs a re-weighting by multiplying the attention map $\mathbf{A}$, to obtain the output of Transformer self-Attention. In this way, patch tokens are reconstructed by leveraging interaction with each other. Formally, self-Attention operation $\mathrm{SA}(.)$ can be formulated as follows:

$$\begin{aligned} \mathbf{Q}, \mathbf{K}, \mathbf{V} &= \hat{\mathbf{S}}\mathbf{W}_q, \hat{\mathbf{S}}\mathbf{W}_k, \hat{\mathbf{S}}\mathbf{W}_v \\ \mathbf{A} &= \mathrm{Softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}})/\sqrt{d} \\ \mathrm{SA}(\hat{\mathbf{S}}) &= \mathbf{A}\mathbf{V} \end{aligned} \tag{1}$$

where $\hat{\mathbf{S}} \in \mathbb{R}^{\hat{N} \times d}$ denotes an 2-dimensional input token sequence, and W_q, W_k, $W_v \in \mathbb{R}^{d \times d'}$ denote three learnable parameter matrices of size $d \times d'$. In the multi-head setting, we let $d' = d/n$, where n indicates the number of attention heads. The function $\mathrm{Softmax}(\cdot)$ denotes the softmax operation for each row. And the scaling operation in Eqn.(1) eliminates the influence from the scale of embedded dimension d'.

In our **Spatio-Temporal Aggregation block** (STA), self-attention operation along time axis and along space axis (i.e. temporal attention and spatial attention) are separately denoted as $\mathrm{SA_t}(\cdot)$ and $\mathrm{SA_s}(\cdot)$. Let $\mathbf{S} \in \mathbb{R}^{\hat{T} \times \hat{N} \times d}$ denote an input spatio-temporal token sequence. Formally, $\mathrm{SA_t}(\cdot)$ and $\mathrm{SA_s}(\cdot)$ can be written as:

$$\begin{aligned} \mathrm{SA_t}(\mathbf{S}) &= \mathrm{SA}(\mathrm{Concat}(\mathbf{S}_{:,0}, ..., \mathbf{S}_{:,\mathrm{n}}, ..., \mathbf{S}_{:,\mathrm{N}-1})) \\ \mathrm{SA_s}(\mathbf{S}) &= \mathrm{SA}(\mathrm{Concat}(\mathbf{S}_{0,:}, ..., \mathbf{S}_{\mathrm{t},:}, ..., \mathbf{S}_{\mathrm{T}-1,:})) \end{aligned} \tag{2}$$

where T indicates the total number of frames in video clip, N is the total spatial position index, and $\mathrm{Concat}(\cdot)$ denotes the concatenation operation in the split dimension, *e.g.*, the spatial position dimension in Eqn.(2).

Given $\mathrm{SA_t}(.)$ and $\mathrm{SA_s}(.)$, the STA block consecutively integrates these two self-attention modules to extract spatial-temporal features. As illustrated in Fig. 3, STA further extracts discriminative information from patch tokens to the class token through spatial attention $\mathrm{SA_s}(\cdot)$, which can be realized by concatenating the copies of class token to the token sequence of each frame before $\mathrm{SA_s}(.)$, and taking the average of class token copies after $\mathrm{SA_s}(.)$ to further apply the later temporal aggregation. In this way, the general form of STA can be presented as:

$$\begin{aligned} \mathbf{S}' &= \mathbf{S} + \alpha \times \mathrm{SA_t}(\mathrm{LN}(\mathbf{S})) \\ \mathrm{STA}(\mathbf{S}, \mathbf{c}) &= \mathrm{Concat}(\mathbf{S}', \mathbf{c}) \\ &+ \beta \times \mathrm{SA_s}(\mathrm{LN}(\mathrm{Concat}(\mathbf{S}', \mathbf{c}))) \end{aligned} \tag{3}$$

where $\mathrm{LN}(\cdot)$ denotes Layer Normalization [3]. The hyper-parameter α and β are learnable scalar residual weights to balance temporal attention and spatial attention. Compared with the space-time joint attention in [5] and [2], which jointly processes all patches of a video, STA is more computation-efficient by reducing complexity from $\mathbb{O}(T^2N^2)$ to $\mathbb{O}(T^2 + N^2)$. Actually, it avoids operating on a long sequence, whose length always leads to quadratic growth of computational complexity [29], [67].

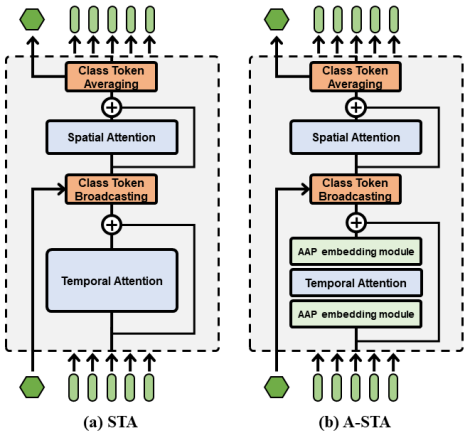


Fig. 3: The detailed comparison between (a) Spatio-Temporal Aggregation block (STA) and (b) Attribution Spatio-Temporal Aggregation block (A-STA). Two additional Attribute-Aware Proxy (AAP) embedding modules are placed into the latter, before and after the temporal attention module. The class token broadcasting operation duplicates the class token for each frame to attend spatial attention within a specific frame. Oppositely, class token averaging calculates the average of all class token copies. Note that the Pre-Norm [78] layers before temporal attention and spatial attention are omitted.

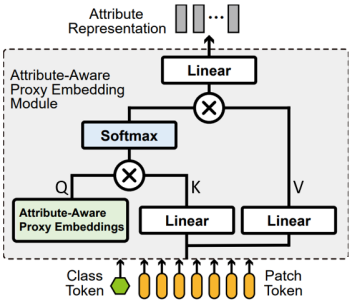


Fig. 4: The detailed module design of the Attribute-Aware Proxy (AAP) embedding module. The Attribute-Aware Proxy Embedding denotes a learnable matrix that is used as the query of the attention operation. For simplicity, this figure only shows the single-head version of the AAP embedding module and the scaling operation before the softmax operation is omitted.

C. Attribute-Aware Proxy Embedding Module

Local patch tokens usually contain rich attributive information, *e.g.,* `glasses`, `umbrellas`, `logos`, and so on. Even if a single attribute is not discriminative enough to recover one's identity, the combinations

of a pedestrian's rich attributes should be discriminative as each attribute eliminates a certain degree of uncertainty. Rather than directly aggregating into a "coarse" class token, we introduce the Attribute-Aware Proxy (AAP) embedding module to directly extract attribute features from a single-frame or multi-frame patch token sequence. Practically, AAP embeddings are formed by a learnable matrix with anisotropic initialization for the richness of learned attributes. It can be considered as the "attribute bank" to serve as the query of the attention operation to match with the feature representations of the input patch tokens. Specifically, AAP embeddings interact with the keys of the patch token sequence. Finally, the resulting attention map is used to re-weight the value, generating the attribute representations of the specific video clip with the same dimension of AAPs. Formally, an AAP embedding module can be written as follows,

$$\begin{aligned} \mathbf{Q},\ \mathbf{K},\ \mathbf{V} &= \mathbf{P}_Q,\ \mathbf{SW}_k,\ \mathbf{SW}_v, \\ \mathrm{AAP}(\mathbf{S}) &= \frac{\mathrm{Softmax}(\mathbf{Q}\mathbf{K}^{\mathrm{T}})}{\sqrt{d}}\mathbf{V} \end{aligned} \tag{4}$$

here we use the multi-head version of AAP embedding module in practice, which has the same multi-head setting as $\mathrm{SA}(\cdot)$ in Eqn.(1). Note that the spatio-temporal input $\mathbf{S}$ here can also be $\hat{\mathbf{S}} \in \mathbb{R}^{\hat{N}\times d}$ for spatial-only use. Compared with $\mathrm{SA}(\cdot)$, the newly proposed AAP module consider the query $\mathbf{Q}$ in Eqn.(1) as the a set of learnable parameters $\mathbf{P}_Q \in \mathbb{R}^{N_a\times d'}$, where $N_a \ll N$ is a hyper-parameter that indicates the number of AAPs. By controlling N_a, the AAP module could have a manually defined capacity, which leads to flexibility for various real applications.

As shown in Fig. 2, both **Stage I** and **Stage III** employ the proposed AAP embedding modules. Specifically, in **Stage I**, the proposed AAP module is firstly used to generate attribute representations from a multi-frame sequence of patch tokens $\mathbf{S} \in \mathbb{R}^{\hat{T}\times\hat{N}\times d}$ for similarity measurement. Although we do not have any attribute-level annotations, we hope the AAP module can automatically learn a rich set of implicit attributes from the entire training dataset, while these resultant attribute representations could also present discriminative power complementary to ID-only representations. To achieve this goal, the ID-level supervision signal is first imposed on the combination of learned attribute representations to constrain its discriminative power. In addition, we initialize the AAPs with anisotropic distributions to capture diverse implicit attribute representations. In practice, we surprisingly find that such anisotropy can maintain after the model training, which means such optimized AAP could respond to a set of differentiated attributes. Moreover, the number of AAPs can be relatively large compared with the class token to cover rich attribute information. In this sense, both the richness and diversity of learned implicit attributes can be guaranteed.

In **Stage III**, we further insert two intra-frame AAP embedding modules before and after the temporal attention of each STA to conduct attribute-aware temporal interaction. Such a modified STA block is named A-STA, which is illustrated in Fig. 3. In A-STA, semantic-related attributes in different frames experience inter-frame interaction to model their temporal relations. In the end, after temporal attention, we set N_a equal to N for the second AAP embedding module so that it has N tokens as output to keep the input-output consistency.

D. Identity-Aware Proxy Embedding Module

Extracting discriminative identity representation is also crucial for video-based re-ID. To this end, the Identity-Aware Proxy (IAP) embedding module is proposed for effective and efficient discriminative representation generation. In previous works, joint space-time attention has shown promising results [2], [5], as it accelerates information aggregation by applying self-attention over spatial and temporal dimensions jointly. However, the quadratic computational overheads limit its applicability. The IAP embedding module is proposed to address such an issue, which performs joint space-time attention with high efficiency while maintaining the discrimination of the identity feature representation.

The IAP module contains a set of identity prototypes, which are presented as two learnable matrics. In practice, we exploit them to replace the keys $\{p_K^i\}_{i=1}^M \in \mathbf{P}_K$ and values $\{p_V^i\}_{i=1}^M \in \mathbf{P}_V$ of the attention operation. Both $\mathbf{P}_K, \mathbf{P}_V \in \mathbb{R}^{M\times d'}$, where $M \in \mathbb{N}^+$ denotes the number of identity prototypes and determine the capacity of the IAP module (usually $M \ll N$). As shown in Fig. 5, an attention map $\mathbf{A} \in \mathbb{R}^{M\times N}$ is first calculated to present the affinity between prototype-patch pairs. Thus each element in $\mathbf{A}$ reflects how close a patch token is to a specific identity prototype. Then this attention map is sparsified by successively applying an L1 normalization and softmax normalization along M and N, respectively. At last, the class token $\mathbf{c}$, *i.e.* the first row of $\mathbf{V}$, is updated by the multiplication of $\mathbf{V}$ and $\mathbf{A}$. Such an operation aggregates the most discriminative identity features from the entire patch token sequence. Formally, given the spatio-temporal token sequence $\mathbf{S}$, the output of the IAP module can be calculated as follows:

$$\begin{aligned} \mathbf{Q},\ \mathbf{K},\ \mathbf{V} &= \mathbf{S}\mathbf{W}_q, \mathbf{P}_K, \mathbf{P}_V \\ \mathbf{A} &= \frac{\mathrm{Softmax}(\mathrm{L1Norm}(\mathbf{Q}\mathbf{K}^{\mathrm{T}}))}{\sqrt{d}} \\ \mathrm{IAP}(\mathbf{S}) &= \mathbf{A}\mathbf{V} \end{aligned} \tag{5}$$

where $\mathbf{K}$ and $\mathbf{V}$ are not conditioned on input $\mathbf{S}$ but are learnable parameters. Here we insert an L1 normalization layer before the softmax operation in Eqn. (5), resulting in double normalization [28], [29]. Such a scheme performs patch token re-coding to reduce the noise of patch representations, leading to robust identification results. Specifically, the learnable

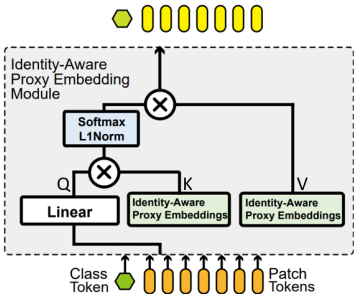


Fig. 5: The detailed module design of the Identity-Aware Proxy (IAP) embedding module. The IAP embedding denotes the learnable matrix used to calculate the key or value of the attention operation. Here we only show the single-head version of the IAP embedding module and omit the scaling operation. In such a scheme, The output token sequence can be considered as reconstruction by a group of IAPs, which tend to reserve the most discriminative identity features.

matrix $\mathbf{P}_K$ matches the input tokens through the double normalization operation to generate the affinity map $\mathbf{A}$. Then these input tokens are thereupon re-coded through a projection of $\mathbf{P}_V$ along $\mathbf{A}$. Since the numbers of learnable vectors in $\mathbf{P}_K$ and $\mathbf{P}_V$ are much smaller than the number of input tokens, the above operation has been able to represent each token in a more compact space (*i.e.* linear combination of the vectors in $\mathbf{P}_V$), effectively suppressing irrelevant information for re-ID. Moreover, $\mathrm{IAP}(\cdot)$ has $O(N)$ computational complexity since the number of identity prototypes M is fixed and is usually much less than the total number of patch tokens of a specific video tracklet (*e.g.,* 64 in our experiments). So, the proposed IAP embedding module allows all spatio-temporal patch tokens to be processed in parallel for effective and efficient feature extraction.

E. Temporal Patch Shuffling

To improve the robustness of the model, we propose a novel data augmentation scheme termed Temporal Patch Shuffling (TPS). Suppose we have one patch sequences $\mathbf{R}_i = \{\mathbf{r}_{i1}, ..., \mathbf{r}_{it}, ..., \mathbf{r}_{iT}\}$ from the same video clip, where the sub-index i denotes specific spatial locations. As shown in Fig. 6, the proposed TPS randomly permutes the patch tokens in $\mathbf{R}_i$ and refill the shuffled sequence $\hat{\mathbf{R}}_i$ to form the new video clip for training. As illustrated in Fig. 6, we could simultaneously select multiple spatial regions in one video clip for shuffling. While in the inference phase, the original video clip is directly fed into the model for identification. TPS brings firm appearance shifts and motion changes from which the network learns to extract generalizable and invariant visual clues. In addition, the scale of available training data can be greatly extended based on such a scheme, which helps to prevent the network from overfitting.

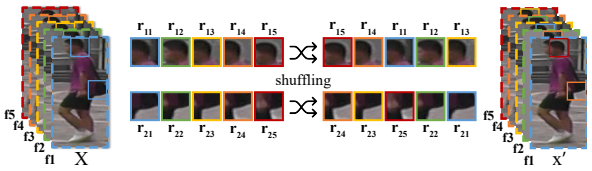


Fig. 6: Visualization of Temporal Patch Shuffling (TPS). $\mathbf{f_t}$ represents t^{th} frame, $\mathbf{r}_{\mathrm{it}}$ the patch in spatial position i and t^{th} frame. TPS is a built-in data augmentation scheme that randomizes the order of a patch sequence sampled from spatial position i. As a result, for example, the patch in the red box is transferred from the 5^{th} frame to the 1^{st} frame.

In our experiments, we treat TPS as a plug-and-play operation and implement it at the stem of the network to promote the entire network for the best performance. The following section will conduct ablation studies to explore where to insert TPS and to what extent TPS should be for optimal training results.

IV. Experiment

A. Datasets and evaluation protocols

In this paper, we evaluate our proposed MSTAT on three widely-used video-based person re-ID benchmarks: iLIDS-VID [68], DukeMTMC-VideoReID (DukeV) [58], and MARS [102].

1) iLIDS-VID [68] is comprised of 600 video tracklets of 300 persons captured from two cameras. In these video tracklets, frame numbers range from 23 and 192. The test set shares 150 identities with the training set.

2) DukeMTMC-VideoReID [58] is a large-scale video-based benchmark which contains $4,832$ videos sharing $1,404$ identities. In the following sections, we use the abbreviation "DukeV" for the DukeMTMC-VideoReID dataset. The video sequences in the DukeV dataset are commonly longer than videos in other datasets, which contain 168 frames on average.

3) MARS [102] is one of the largest video re-ID benchmarks which collects $1,261$ identities existing in around $20,000$ video tracklets captured by 6 cameras. Frames within a video tracklet are relatively more misaligned since they are obtained by a DPM detector [25] and a GMMCP tracker [20] rather than hand drawing. Furthermore, around $3,200$ distractor tracklets are mixed into the dataset to simulate real-world scenarios.

For evaluation on MARS and DukeV datasets, we use two metrics: the Cumulative Matching Characteristic (CMC) curves [6] and mean Average Precision (mAP) following previous works [14], [50], [94], [99]. However, in the gallery set of iLIDS-VID, there is merely one correct match for each query. For this benchmark, only cumulative accuracy is reported.

Method	Source	Backbone	MARS			Duke-V			iLIDS-VID	
			Rank-1	Rank-5	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
SCAN [94]	TIP19	Pure-CNN	87.2	95.2	77.2	-	-	-	88.0	96.7
VRSTC [37]	CVPR19	Pure-CNN	-	89.8	85.1	96.9	-	96.2	86.6	-
M3D [37]	AAAI19	Pure-CNN	-	-	-	96.9	-	96.2	74.0	94.3
MG-RAFA [99]	CVPR20	Pure-CNN	88.8	97.0	85.9	-	-	-	88.6	98.0
AFA [14]	ECCV20	Pure-CNN	90.2	96.6	82.9	-	-	-	88.5	96.8
AP3D [27]	ECCV20	Pure-CNN	90.7	-	85.6	97.2	-	96.1	88.7	-
TCLNet [14]	ECCV20	Pure-CNN	89.8	-	85.1	96.9	-	96.2	86.6	-
A3D [13]	TIP20	Pure-CNN	86.3	95.5	80.4	-	-	-	86.7	98.6
GRL [51]	CVPR21	Pure-CNN	90.4	96.7	84.8	95.0	98.7	93.8	90.4	98.3
STRF [1]	ICCV21	Pure-CNN	90.3	-	86.1	97.4	-	96.4	89.3	-
Fang *et al.* [23]	WACV21	Pure-CNN	87.9	97.2	83.2	-	-	-	88.6	98.6
TMT [50]	Arxiv21	CNN-Transformer	91.2	97.3	85.8	-	-	-	91.3	98.6
Liu *et al.* [46]	CVPR21*	CNN-Transformer	91.3	-	**86.5**	96.7	-	96.2	-	-
STT [95]	Arxiv21	CNN-Transformer	88.7	-	86.3	**97.6**	-	**97.4**	87.5	95.0
ASANet [8]	TCSVT22	Pure-CNN	91.1	97.0	86.0	**97.6**	**99.9**	97.1	-	-
MSTAT(ours)	-	Pure-Transformer	**91.8**	**97.4**	85.3	97.4	99.3	96.4	**93.3**	**99.3**

TABLE I: Result comparison with state-of-the-art video-based person re-ID methods on MARS, DukeMTMC-VideoReID, and iLIDS-VID. * denotes the workshop of the conference.

B. Implementation details

Our proposed MSTAT framework is built based on Pytorch toolbox [57]. In our experiments, it is running on a single NVIDIA A100 GPU (40G memory). We resize each video frame to 224×112 for the above benchmarks. Typical data augmentation schemes are involved in training, including horizontal flipping, random cropping, and random erasing. For all stages, STA modules are pretrained on an action recognition dataset, K600 [7], while other aforementioned modules are randomly initialized.

In the training phase, if not specified, we sample $L = 8$ frames each time for a video tracklet and set the batch size as 24. In each mini-batch, we randomly sample two video tracklets from different cameras for each person. We supervise the network by cross-entropy loss with label smoothing [63] associated with widely used BatchHard triplet loss [34]. Specifically, we impose supervision signals separately on the concatenated attribute representation from the AAP embedding module in **Stage I**, the output class tokens from **Stage II**, and **Stage III**. The learning rate is initially set to 1e-3, which would be multiplied by 0.75 after every 25 epochs. The entire network is updated by an SGD optimizer in which the weight decay and Nesterov momentum are set to 5×10^5 and 0.9, respectively.

In the test phase, following [27], [95], we randomly sample 32 frames as a sequence from each original tracklet in either query or gallery. For each sequence, The attribute representation from **Stage I**, the output class tokens from stage **Stage II** and **Stage III** are concatenated as the overall representation. Following the widely-used protocol, we compute the cosine similarity between each query-gallery pair using their overall representations. Then, the CMC curves and the mAP can be calculated based on the predicted ranking list and the ground truth identity of each query. Note that we do not use any re-ranking technique.

C. Compared with the state of the arts

In Table I, we make a comparison on three benchmark datasets between our method and video-based person re-ID methods from 2019 to 2021, including M3D [37], GRL [51], STRF [1], Fang *et al.* [23], TMT [50], Liu *et al.* [46], ASANet [8]. According to their backbones, these re-ID methods can be roughly divided into the following types: Pure-CNN, CNN-Transformer Hybrid, and Pure-Transformer methods.

In real-world applications, rank-1 accuracy [6] reflects what extent a method can find the most confident positive sample [85], and relatively high rank-1 accuracy can save time in confirmation. As the first method based on Pure-Transformer for video-based re-ID so far, we achieve state-of-the-art results in rank-1 accuracy on three benchmarks. Our approach especially attains rank-1 accuracy of 91.8% and rank-5 accuracy of 97.4% on the largest-scale benchmark, MARS. It is noteworthy that our MSTAT outperforms the best pure CNN-based methods using ID annotations only by a margin of 1.1% and a CNN-Transformer hybrid method, TMT, by 0.6% in MARS rank-1 accuracy.

Compared to our proposed method, TCLNet [14] explicitly captures complementary features over different frames, and GRL [51] devises a guiding mechanism for reciprocating feature learning. However, the designed modules in these methods commonly take as input the deep spatial feature maps extracted by a CNN backbone (*e.g.* ResNet50) that may overlook attribute-associated or identity-associated information without explicit modeling. Similar to ours, TMT [50] and M3D [1] process video tracklets in multiple views to extract and fuse multi-view features. Notably, in all stages of MSTAT, intermediate features are spatio-temporal and can be iteratively updated to capture spatio-temporal cues with different emphases. ASANet [8] exploits explicit ID-relevant attributes (*e.g.*, *gender, clothes, and hair*) and ID-irrelevant attributes (*e.g.*, *pose and motion*) on a multi-branch net-

Method	Test Protocol	Rank-1	Rank-5	mAP
MSTAT	Stage I	89.2	96.7	82.4
	Stage II	89.2	96.5	83.0
	Stage III	89.8	96.5	83.0
	Stage I & II	91.2	97.3	85.0
	Stage I & III	90.5	97.2	83.9
	Stage II & III	90.6	96.9	84.6
	Stage I, II, & III	**91.8**	**97.4**	**85.3**

TABLE II: Ablation study on three stages of MSTAT on MARS. Test Protocol means the final feature representation used for similarity measurement. The network architecture and training hyper-parameter setting remain the same for each experiment.

work. Despite the performance growth, the demand for attribute annotations may limit its applications in large-scale scenarios. In comparison with existing methods, our method aggregates spatio-temporal information in a unified manner and explicitly capitalizes on implicit attribute information to improve recognizability under challenging scenarios. Conclusively, our method achieves the state-of-the-art performance of 91.8% and 93.3% rank-1 accuracy, respectively, on MARS and iLIDS-VID.

D. Effectiveness of Multi-Stage Framework Architecture

To evaluate the effectiveness of the three stages in our proposed MSTAT, we carry out a series of ablation experiments whose results are displayed in Table II. After the three stages are jointly trained, we first separately evaluate each stage using its output feature representation. Then, we concatenate two or more stages to evaluate whether each is effective.

For three single stages, each has rank-1 accuracy ranging from 89.2 to 89.8. However, their combinations result in a significant increase of over 0.8%. Remarkably, while **Stage I** and **Stage II** secure only 89.2 rank-1 accuracy, their integration attains up to 91.2%, surpassing them by a 2% margin. One can attribute such a result to their emphases: one stage on attribute-associated features and the other on identity-associated features. Eventually, when all three stages are used, MSTAT reaches a 91.8% rank-1 accuracy, higher than all two-stage combinations. Overall, these experiments demonstrate that the three stages have different preferences toward features and can complement each other by simple concatenation.

E. Effectiveness of Key Components

To demonstrate the effectiveness of our proposed MSTAT, we conduct a range of ablative experiments on the largest public benchmark MARS.

1) Effectiveness of Attribute-Aware Proxy Embedding Module: As shown in Fig. 7, we evaluate MSTAT with different AAP numbers (*i.e.* N_a in Sec. III-C) in the AAP embedding module in the last layer of **Stage I**. The figure reveals that 24 proxies are optimal for

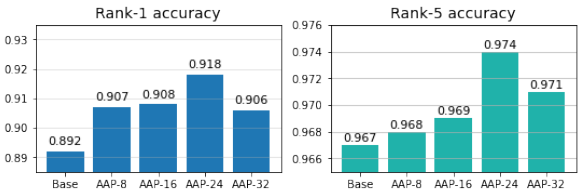


Fig. 7: Ablation study on the attribute-aware proxy (AAP) embedding module for attribute extraction in MARS. "Base" is the network without attribute extraction using AAP in training and testing. AAP-**k** indicates the network where the AAP embedding module in **Stage I** has **k** AAPs.

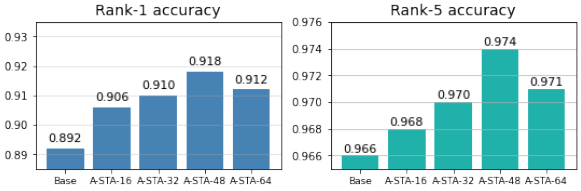


Fig. 8: Ablation study on A-STA. "Base" is the network that consists of STA only. A-STA-**k** represents the network in which **Stage III** is equipped with A-STA layers each of **k** AAPs.

attributive information extraction as it attains the best performance in terms of rank-1 and rank-5 accuracy. In contrast to the baseline, MSTAT has seen over 2% growth in rank-1 accuracy and around 1% in rank-5 accuracy. However, a redundant or insufficient number of AAPs may cause a minor performance drop since they may pay attention to noisy or useless attributes. In summary, the AAP embedding module for clue extraction gives a boost to the performance in rank-1 and rank-5 accuracy, with negligible computational overhead.

Attribute-Aware Proxy (AAP) embedding modules are also used for A-STA, a variant of STA for attribute-aware temporal feature fusion in **Stage III**. As shown in Fig. 8, we conduct a series of experiments to explore whether A-STA is effective and how many AAPs for A-STA are appropriate (also corresponding to N_a in III-C)). The experiment results reveal that the baseline model fails to reach 90% rank-1 accuracy or 97% rank-5 accuracy. As the number of AAPs increases, these two metrics grow to 91.8% and 97.4%.

Therefore, we can attribute the performance soar to A-STA, allowing for attribute-aware temporal interaction. A-STA offers a different viewpoint from that of **Stage II** on videos. Moreover, due to the redundancy of temporal information in many video re-ID scenarios discussed in [14], A-STA with too many AAPs incurs meaningless attributes. This can be why the performance descends once A-STA has too many AAPs.

In conclusion, our proposed AAP embedding module can be used for: (1) the extraction of informative attributes as plugged into any Transformer layer and (2) attribute-aware temporal interaction when a tempo-

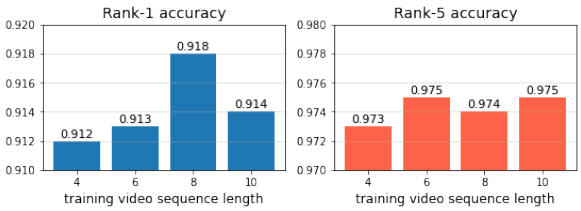


Fig. 9: Study on the effect of training video sequence length on MARS.

Method	Position	Rank-1	Rank-5
w/o IAP embedding module	-	88.2	96.4
w/ IAP embedding module	Stage II	91.0	97.0
	Stage III	90.4	97.0
	Stage II&III	**91.8**	**97.4**

TABLE III: Ablation study on the IAP embedding module. Stage II and Stage III in this table means that an IAP embedding module is appended to the last layer of Stage II and Stage III respectively. This table shows that the IAP embedding module brings improvements to every single stage. When it is placed on both two stages, MSTAT shows the best performance.

ral attention module is sandwiched between two. Both of the two functionalities cause a significant increase in performance, demonstrating their effectiveness.

2) Effectiveness of Identity-Aware Proxy Embedding Module: In Table III, MSTAT that discards IAP embedding modules leads to only 88.2% rank-1 accuracy and 96.4% rank-5 accuracy. However, it boosts rank-1 performance by 2.8% or 2.2% by taking the place of STA in **Stage II** or **Stage III**. Finally, IAP embedding modules in the last layers in both **Stage II** and **Stage III** further improve 0.8% rank-1 accuracy and 0.4% rank-5 accuracy. The IAP embedding module's ablation results demonstrate its ability to generate discriminative representations efficiently. Intuitively, we place the IAP embedding module only in the last few depths because it may discard non-discriminative features that should be preserved in shallow layers.

3) Effectiveness of Temporal Patch Shuffling: To evaluate the effectiveness of Temporal Patch Shuffling (TPS), we assign different probabilities to implement TPS for each training video sample. Note that in the following experiments, the number of spatial positions to shuffle is set to 5 if we implement TPS on this sample. As shown in Table IV, 20% probability provides the best result over others, which leads to a growth of 0.3% in rank-1 accuracy. However, the 60% or 80% probability results in a 0.1% or 0.2% rank-1 accuracy drop mainly due to heavy noise. In summary, a proper level of TPS would be an effective data augmentation method for the Transformer for video-based person re-ID. Further, rather than reserving temporal motion (an ordered sequence of patches), TPS stimulates re-identification accuracy by learning temporal coherence from shuffled patch tokens.

Methods	Prob.	Rank-1	Rank-5	mAP
MSTAT w/o TPS	0%	91.5	**97.5**	85.2
MSTAT w/ TPS	20%	**91.8**	97.4	**85.3**
	40%	91.7	97.3	85.2
	60%	91.4	**97.5**	85.1
	80%	91.3	97.1	85.1

TABLE IV: Ablation study on Temporal Patch Shuffling. The table shows that the proper level of shuffling can bring slight improvement. However, it may degrade the learning while the shuffling degree becomes increasingly overwhelming.

F. Effect of video sequence length

To investigate how temporal noise influences the training of MSTAT, we conduct experiments on videos with varied lengths. In Fig. 9, experiments provide length-varying video tracklets for training, while all experiments are implemented under the identical evaluation setting with a fixed video length of 32. All experiments shut down until the loss stops decreasing for ten epochs.

On the one hand, rank-1 accuracy shows an upward trend as temporal noise gradually decreases, reaching a peak at 8. On the other hand, temporal noise shows no apparent correlation with rank-5 accuracy and mAP. These results show that our model gains up to 0.6% rank-1 accuracy through learning better temporal features from data. However, rank-5 accuracy and mAP benefit little from noise reduction, from which we can speculate that in most cases in video re-ID, learning temporal features is less important than learning appearance features as they only account for 0.6% of rank-1 and 0.2% of rank-5 accuracy. Similar results can be found in [50].

G. Comparison among metric learning methods

Metric learning aims to regularize the sample distribution on feature space. Usually, metric learning losses constrain the compactness of intra-class distribution and sparsity of the overall distribution. To explore which strategy cooperates with our framework better, we compare a range of classic metric learning loss functions on iLIDS-VID, as shown in Table V. Note that these losses are scaled to the same magnitude to ensure fairness. Significantly, OIM loss [77] and BatchHard triplet loss [34], widely used in re-ID, outperform Arcface [21] and SphereFace [49] losses by a large margin since the latter two loss functions suffer from untimely overfitting in our experiments.

H. Visualization

To better understand how the proposed framework works, we conduct visualization on the AAP embedding module. In Fig. 10, we show the diversity of implicit attributes by the similarity matrix of 24 AAPs. This figure implies that AAPs are anisotropic, covering

Metric learning loss	Rank-1	Rank-5
w/o Metric learning	66.0	90.0
Arcface [21]	73.3	90.7
SphereFace [49]	66.7	89.3
OIM [77]	89.3	98.3
BatchHard* [34]	**93.3**	**99.3**

TABLE V: Comparison among metric learning loss functions on iLIDS-VID, where * denotes the method used in our implementation. For Arcface and SphereFace, we test three margins and report the best result: (1) by default, (2) 20% larger than the default, (3) 20% smaller than the default.

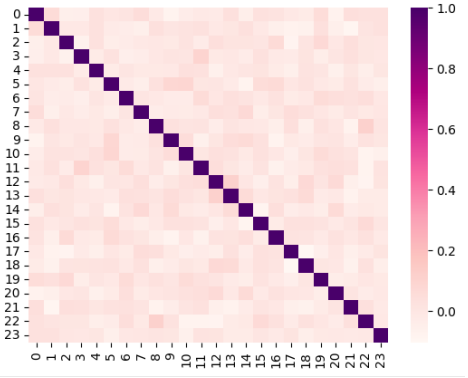


Fig. 10: Visualization of the similarity matrix of attribute-aware proxies trained on MARS. The maximal similarity between all pairs is around 0.2, demonstrating that AAPs learn to capture diverse attributes.

different attribute features that appear in the given training dataset.

Specifically, as shown in Fig. 11, we randomly select two pedestrians' tracklets. Attention map visualization is adopted as a sign of each AAP's concentration. In practice, we process the raw attention maps first by several average filters and then by thresholding to deliver smooth visual effects instead of grid-like maps. In these heap maps, the brighter color denotes the higher attention value. Despite the absence of attribute-level supervision, Fig.11 shows that some AAPs learn to pay attention to a local region with special meanings as an identity cue. For example, the AAP in white color in video clips (a) automatically learns to cover the logo in the T-shirt, while the one in (b) captures the head of the woman.

Moreover, we display the t-SNE visualization result on iLIDS-VID in Fig. 12. It only contains the first 1/3 of the IDs in the test set for a better visual effect. We also provide the corresponding quantitative evaluation results in Table VI measured by the normalized averaged intra-class distance and the minimum inter-class distance (0-2) on the entire test set. As a result, MSTAT drops the average intra-class distance from 0.4572 of the baseline to 0.4410 and enlarges the minimum inter-class distance from 0.4704 to 0.5012. Further, to eliminate the influence of accuracy, we measure the intra-class distance between correctly

Methods	Intra↓	Intra*↓	Inter↑	Rank-1↑
Baseline [5]	0.4572	0.4495	0.4704	0.873
MSTAT w/o attr.	0.4517	0.4469	0.4644	0.913
MSTAT (ours)	**0.4410**	**0.4389**	**0.5012**	**0.933**

TABLE VI: Quantitative evaluation on iLIDS-VID. "Intra" denotes the averaged normalized intra-class distance, and "Inter" is the minimum inter-class distance. Here, * means that the metric is computed on samples with the correct rank-1 match.

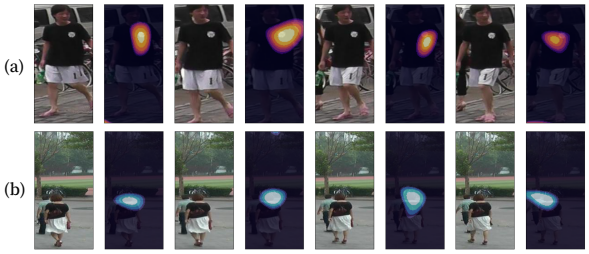


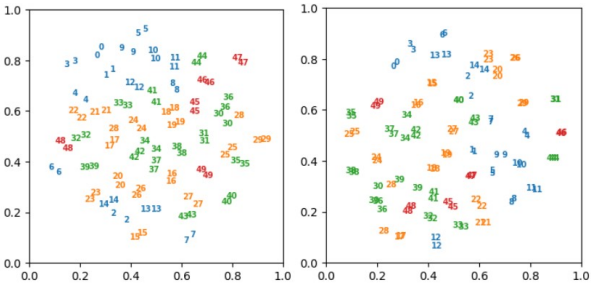
Fig. 11: Visualization of attribute-aware proxies for two different pedestrians on MARS. Attention heat maps of four consecutive frames from the AAP embedding module on Stage I are displayed.

matched samples, from which we witness a similar result. These results explain why MSTAT's t-SNE visualization seems sparser.

V. Conclusion

This paper proposes a novel framework for video-based person re-ID, referred to as Spatial-Temporal Aggregation Transformer (MSTAT). To tackle simultaneous extraction for local attributes and global identity information, MSTAT adopts a multi-stage architecture to extract (1) attribute-associated, (2) the identity-associated, and (3) the attribute-identity-associated information from video clips, with all layers inherited from the vanilla Transformer. Further, for reserving informative attribute features and aggregating discriminative identity features, we introduce two proxy embedding modules (Attribute-Aware Proxy embedding module and Identity-Aware Proxy embedding module) into different stages. In addition, a patch-based data augmentation scheme, Temporal Patch Shuffling, is proposed to force the network to learn invariance to appearance shifts while enriching training data. Massive experiments show that MSTAT can extract attribute-aware features consistent across frames while reserving discriminative global identity information on different stages to attain high performance. Finally, MSTAT outperforms most existing state-of-the-arts on three public video-based re-ID benchmarks.

Future work may focus on mining the hard instances or local informative attribute locations to conduct contrastive learning to promote the model's accuracy further. Moreover, leveraging more unlabeled and multi-



(a) Baseline (b) MSTAT

Fig. 12: T-SNE Visualization of the iLIDS-VID test set. The numbers on the plots indicate person IDs. MSTAT shows an increase in intra-class compactness and the minimum inter-class distance over the entire test set compared to the baseline.

modal data to improve the model's effectiveness is also a potential research direction.

ACKNOWLEDGMENT

The work is supported in part by the Young Scientists Fund of the National Natural Science Foundation of China under grant No. 62106154, by National Key R&D Program of China under Grant No. 2021ZD0111600, by Natural Science Foundation of Guangdong Province, China (General Program) under grant No.2022A1515011524, by Guangdong Basic and Applied Basic Research Foundation under Grant No. 2017A030312006, by CCF-Tencent Open Fund, by Shenzhen Science and Technology Program ZDSYS20211021111415025, and by the Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese Univeristy of Hong Kong (Shenzhen).

REFERENCES

[1] A. Aich, M. Zheng, S. Karanam, T. Chen, A. K. Roy-Chowdhury, and Z. Wu. Spatio-temporal representation factorization for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 152–162, 2021.

[2] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021.

[3] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] S. Bak and P. Carr. One-shot metric learning for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2990–2999, 2017.

[5] G. Bertasius, H. Wang, and L. Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[6] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior. The relation between the roc curve and the cmc. In *Fourth IEEE workshop on automatic identification advanced technologies (AutoID'05)*, pages 15–20. IEEE, 2005.

[7] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

[8] T. Chai, Z. Chen, A. Li, J. Chen, X. Mei, and Y. Wang. Video person re-identification using attribute-enhanced features. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[9] Z. Chang, Z. Yang, Y. Chen, Q. Zhou, and S. Zheng. Seq-masks: Bridging the gap between appearance and gait modeling for video-based person re-identification. In *2021 International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2021.

[10] C. Chen, M. Ye, M. Qi, J. Wu, Y. Liu, and J. Jiang. Saliency and granularity: Discovering temporal coherence for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[11] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1169–1178, 2018.

[12] G. Chen, J. Lu, M. Yang, and J. Zhou. Spatial-temporal attention-aware learning for video-based person re-identification. *IEEE Transactions on Image Processing*, 28(9):4192–4205, 2019.

[13] G. Chen, J. Lu, M. Yang, and J. Zhou. Learning recurrent 3d attention for video-based person re-identification. *IEEE Transactions on Image Processing*, 29:6963–6976, 2020.

[14] G. Chen, Y. Rao, J. Lu, and J. Zhou. Temporal coherence or temporal motion: Which is more critical for video-based person re-identification? In *European Conference on Computer Vision*, pages 660–676. Springer, 2020.

[15] T. Chen, L. Lin, R. Chen, X. Hui, and H. Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 44(03):1371–1384, 2022.

[16] T. Chen, T. Pu, H. Wu, Y. Xie, L. Liu, and L. Lin. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[17] X. Chen, J. Xu, J. Xu, and S. Gao. Oh-former: Omni-relational high-order transformer for person re-identification. *arXiv preprint arXiv:2109.11159*, 2021.

[18] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016.

[19] D. Chung, K. Tahboub, and E. J. Delp. A two stream siamese convolutional neural network for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 1983–1991, 2017.

[20] A. Dehghan, S. Modiri Assari, and M. Shah. Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4091–4099, 2015.

[21] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[22] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[23] P. Fang, P. Ji, L. Petersson, and M. Harandi. Set augmented triplet loss for video person re-identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 464–473, 2021.

[24] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2360–2367. IEEE, 2010.

[25] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[26] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[27] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen. Appearance-preserving 3d convolution for video-based person re-identification. In *European Conference on Computer*

Vision, pages 228–243. Springer, 2020.

[28] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

[29] M.-H. Guo, Z.-N. Liu, T.-J. Mu, and S.-M. Hu. Beyond self-attention: External attention using two linear layers for visual tasks. *arXiv preprint arXiv:2105.02358*, 2021.

[30] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang. Transformer in transformer. *arXiv preprint arXiv:2103.00112*, 2021.

[31] X. Hao, S. Zhao, M. Ye, and J. Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16403–16412, 2021.

[32] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang. Transreid: Transformer-based object re-identification. *arXiv preprint arXiv:2102.04378*, 2021.

[33] T. He, X. Jin, X. Shen, J. Huang, Z. Chen, and X.-S. Hua. Dense interaction learning for video-based person re-identification. *arXiv preprint arXiv:2103.09013*, 2021.

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

[35] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[36] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen. Temporal complementary learning for video person re-identification. In *European conference on computer vision*, pages 388–405. Springer, 2020.

[37] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen. Vrstc: Occlusion-free video person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2019.

[38] S. Karanam, Y. Li, and R. J. Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE international conference on computer vision*, pages 4516–4524, 2015.

[39] J. Li, S. Zhang, and T. Huang. Multi-scale 3d convolution network for video based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8618–8625, 2019.

[40] M. Li, X. Zhu, and S. Gong. Unsupervised person re-identification by deep learning tracklet association. In *Proceedings of the European conference on computer vision (ECCV)*, pages 737–753, 2018.

[41] S. Li, S. Bak, P. Carr, and X. Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 369–378, 2018.

[42] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014.

[43] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015.

[44] X. Liao, L. He, Z. Yang, and C. Zhang. Video-based person re-identification via 3d convolutional networks and non-local attention. In *Asian Conference on Computer Vision*, pages 620–634. Springer, 2018.

[45] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person re-identification: What features are important? In *European Conference on Computer Vision*, pages 391–401. Springer, 2012.

[46] C.-T. Liu, J.-C. Chen, C.-S. Chen, and S.-Y. Chien. Video-based person re-identification without bells and whistles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1491–1500, 2021.

[47] C.-T. Liu, C.-W. Wu, Y.-C. F. Wang, and S.-Y. Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. *arXiv preprint arXiv:1908.01683*, 2019.

[48] H. Liu, Z. Jie, K. Jayashree, M. Qi, J. Jiang, S. Yan, and J. Feng. Video-based person re-identification with accumulative motion context. *IEEE transactions on circuits and systems for video technology*, 28(10):2788–2802, 2017.

[49] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[50] X. Liu, P. Zhang, C. Yu, H. Lu, X. Qian, and X. Yang. A video is worth three views: Trigeminal transformers for video-based person re-identification. *arXiv preprint arXiv:2104.01745*, 2021.

[51] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang. Watching you: Global-guided reciprocal learning for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13334–13343, 2021.

[52] Y. Liu, Z. Yuan, W. Zhou, and H. Li. Spatial and temporal mutual promotion for video-based person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8786–8793, 2019.

[53] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[54] N. McLaughlin, J. M. Del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2016.

[55] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[56] A. Nambiar, A. Bernardino, and J. C. Nascimento. Gait-based person re-identification: A survey. *ACM Computing Surveys (CSUR)*, 52(2):1–34, 2019.

[57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[58] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[59] Y. Shi, Z. Wei, H. Ling, Z. Wang, J. Shen, and P. Li. Person retrieval in surveillance videos via deep attribute mining and reasoning. *IEEE Transactions on Multimedia*, 23:4376–4387, 2020.

[60] W. Song, J. Zheng, Y. Wu, C. Chen, and F. Liu. A two-stage attribute-constraint network for video-based person re-identification. *IEEE Access*, 7:8508–8518, 2019.

[61] A. Subramaniam, A. Nambiar, and A. Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 562–572, 2019.

[62] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018.

[63] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[64] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *European conference on computer vision*, pages 791–808. Springer, 2016.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[66] C. Wan, Y. Wu, X. Tian, J. Huang, and X.-S. Hua. Concentrated local part discovery with fine-grained part representation for person re-identification. *IEEE Transactions on Multimedia*, 22(6):1605–1618, 2019.

[67] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma. Linformer: Self-attention with linear complexity. 2020.

[68] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *European conference on computer vision*, pages 688–703. Springer, 2014.

[69] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv*

preprint arXiv:2102.12122, 2021.

[70] A. Wu, W.-S. Zheng, and J.-H. Lai. Robust depth-based person re-identification. *IEEE Transactions on Image Processing*, 26(6):2588–2603, 2017.

[71] D. Wu, M. Ye, G. Lin, X. Gao, and J. Shen. Person re-identification by context-aware part attention and multi-head collaborative learning. *IEEE Transactions on Information Forensics and Security*, 17:115–126, 2021.

[72] L. Wu, Y. Wang, J. Gao, and X. Li. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Transactions on Multimedia*, 21(6):1412–1424, 2018.

[73] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W.-S. Zheng. An enhanced deep feature representation for person re-identification. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016.

[74] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang. Exploit the unknown gradually: One-shot video-based person re-identification by stepwise learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2018.

[75] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer. Second-order non-local attention networks for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3760–3769, 2019.

[76] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1249–1258, 2016.

[77] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3415–3424, 2017.

[78] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

[79] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 4733–4742, 2017.

[80] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *European Conference on Computer Vision*, pages 701–716. Springer, 2016.

[81] M. Ye, C. Chen, J. Shen, and L. Shao. Dynamic tri-level relation mining with attentive graph for visible infrared re-identification. *IEEE Transactions on Information Forensics and Security*, 17:386–398, 2021.

[82] M. Ye, X. Lan, Q. Leng, and J. Shen. Cross-modality person re-identification via modality-aware collaborative ensemble learning. *IEEE Transactions on Image Processing*, 29:9387–9399, 2020.

[83] M. Ye, H. Li, B. Du, J. Shen, L. Shao, and S. C. Hoi. Collaborative refining for person re-identification with label noise. *IEEE Transactions on Image Processing*, 31:379–391, 2021.

[84] M. Ye, J. Shen, D. J Crandall, L. Shao, and J. Luo. Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 229–247. Springer, 2020.

[85] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[86] M. Ye, J. Shen, and L. Shao. Visible-infrared person re-identification via homogeneous augmented tri-modal learning. *IEEE Transactions on Information Forensics and Security*, 16:728–739, 2020.

[87] F. Yu, X. Jiang, Y. Gong, S. Zhao, X. Guo, W.-S. Zheng, F. Zheng, and X. Sun. Devil's in the details: Aligning visual clues for conditional embedding in person re-identification. *arXiv preprint arXiv:2009.05250*, 2020.

[88] Z. Yu, Y. Zhao, B. Hong, Z. Jin, J. Huang, D. Cai, X. He, and X.-S. Hua. Apparel-invariant feature learning for person re-identification. *IEEE Transactions on Multimedia*, 2021.

[89] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, F. E. Tay, J. Feng, and S. Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021.

[90] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun. Feature pyramid transformer. In *European Conference on Computer Vision*, pages 323–339. Springer, 2020.

[91] J. Zhang, N. Wang, and L. Zhang. Multi-shot pedestrian re-identification via sequential decision making. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6781–6789, 2018.

[92] L. Zhang, Z. Shi, J. T. Zhou, M.-M. Cheng, Y. Liu, J.-W. Bian, Z. Zeng, and C. Shen. Ordered or orderless: A revisit for video based person re-identification. *IEEE transactions on pattern analysis and machine intelligence*, 43(4):1460–1466, 2020.

[93] P. Zhang, J. Xu, Q. Wu, Y. Huang, and X. Ben. Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild. *IEEE Transactions on Multimedia*, 23:3562–3576, 2020.

[94] R. Zhang, J. Li, H. Sun, Y. Ge, P. Luo, X. Wang, and L. Lin. Scan: Self-and-collaborative attention network for video person re-identification. *IEEE Transactions on Image Processing*, 28(10):4870–4882, 2019.

[95] T. Zhang, L. Wei, L. Xie, Z. Zhuang, Y. Zhang, B. Li, and Q. Tian. Spatiotemporal transformer for video-based person re-identification. *arXiv preprint arXiv:2103.16469*, 2021.

[96] Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, H. Chen, I. Marsic, and J. Tighe. Vidtr: Video transformer without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13577–13587, 2021.

[97] Y. Zhang, S. Zhao, Y. Kang, and J. Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 462–479. Springer, 2022.

[98] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019.

[99] Z. Zhang, C. Lan, W. Zeng, and Z. Chen. Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10407–10416, 2020.

[100] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1077–1085, 2017.

[101] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4913–4922, 2019.

[102] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[103] W.-S. Zheng, S. Gong, and T. Xiang. Towards open-world person re-identification by one-shot group-based verification. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):591–606, 2015.

[104] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019.

[105] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4747–4756, 2017.

Ziyi Tang is now pursuing his Ph.D. degree at Sun Yat-Sen University. Before that, he was a research assistant at The Chinese University of Hong Kong, Shenzhen (CUHK-SZ), China. He received the B.E. degree from South China Agriculture University (SCAU), Guangzhou, China in 2019 and M.S. degree from The University of Southampton, Southampton, U.K. in 2020. He has won top places in data science competitions hosted by Kaggle and Huawei respectively. His research interests include Computer Vision, Vision-Language Joint Modeling, and Casual Inference.

Ruimao Zhang is currently a Research Assistant Professor in the School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-SZ), China. He is also a Research Scientist at Shenzhen Research Institute of Big Data. He received the B.E. and Ph.D. degrees from Sun Yat-sen University, Guangzhou, China in 2011 and 2016, respectively. From 2017 to 2019, he was a Post-doctoral Research Fellow in the Multimedia Lab, The Chinese University of Hong Kong (CUHK), Hong Kong. After that, he joined at SenseTime Research as a Senior Researcher until 2021. His research interests include computer vision, deep learning and related multimedia applications. He has published about 40 peer-reviewed articles in top-tier conferences and journals such as TPAMI, IJCV, ICML, ICLR, CVPR, and ICCV. He has won a number of competitions and awards such as Gold medal in 2017 Youtube 8M Video Classification Challenge, the first place in 2020 AIM Challenge on Learned Image Signal Processing Pipeline. He was rated as Outstanding Reviewer of NeurIPS in 2021. He is a member of IEEE.

Zhanglin Peng is now pursuing her Ph.D. degree with the Department of Computer Science, The University of Hong Kong, Hong Kong, China. She received her B.E. and M.S. degrees from Sun Yat-Sen University, Guangzhou, China in 2013 and 2016, respectively. From 2016 to 2020, she was a researcher at SenseTime Research. Her research interests are computer vision and machine learning.

Jinrui Chen is currently pursuing the B.A. degree in Financial Engineering conferred jointly by the School of Data Science, the School of Science and Engineering, and the School of Management and Economics, The Chinese University of Hong Kong, Shenzhen (CUHK-SZ), China. His research interests include deep learning and financial technology.

Liang Lin (M'09, SM'15) is a Full Professor of computer science at Sun Yat-sen University. He served as the Executive Director and Distinguished Scientist of SenseTime Group from 2016 to 2018, leading the R&D teams for cutting-edge technology transferring. He has authored or co-authored more than 200 papers in leading academic journals and conferences, and his papers have been cited by more than 22,000 times. He is an associate editor of IEEE Trans. Multimedia and IEEE Trans. Neural Networks and Learning Systems, and served as Area Chairs for numerous conferences such as CVPR, ICCV, SIGKDD and AAAI. He is the recipient of numerous awards and honors including Wu Wen-Jun Artificial Intelligence Award, the First Prize of China Society of Image and Graphics, ICCV Best Paper Nomination in 2019, Annual Best Paper Award by Pattern Recognition (Elsevier) in 2018, Best Paper Dimond Award in IEEE ICME 2017, Google Faculty Award in 2012. His supervised PhD students received ACM China Doctoral Dissertation Award, CCF Best Doctoral Dissertation and CAAI Best Doctoral Dissertation. He is a Fellow of IET and IAPR.