A Video Is Worth Three Views: Trigeminal Transformers for Video-Based Person Re-Identification

Xuehu Liu, Pingping Zhang®, Chenyang Yu®, Xuesheng Qian, Xiaoyun Yang, and Huchuan Lu®, Fellow, IEEE

Abstract—Video-based person Re-Identification (Re-ID) is a hot research topic in intelligent transportation systems, which aims to retrieve video sequences of the same person under non-overlapping surveillance cameras. Compared with static images, video sequences contain more visual information from multiple views, such as spatial and temporal views. However, previous Re-ID methods usually focus on single limited views, lacking diverse observations from different views. To capture richer perceptions and extract more comprehensive representations, we propose a novel learning framework named Trigeminal Transformers (TMT) to tackle video-based person Re-ID. More specifically, we first design a View-wise Projector (VP) to jointly transform raw videos from spatial, temporal and spatial-temporal views. In addition, inspired by the great success of Vision Transformers (ViT), we introduce the Transformer structure for information enhancement and aggregation. In our work, three Self-view Transformers (ST) are proposed to exploit the relationships of local features for information enhancement in spatial, temporal and spatial-temporal. Moreover, a Cross-view Transformer (CT) is proposed to aggregate the multi-view features for comprehensive representations. Experimental results indicate that our approach can obtain better performance than some other state-of-the-art approaches on four public Re-ID benchmarks.

Index Terms—Person re-identification, vision transformer, video representation, deep feature aggregation, deeply-supervised learning.

I. INTRODUCTION

Object Re-identification (Re-ID) is a hot research topic in intelligent transportation systems [1], which aims to retrieve target objects across different times and places, such as person Re-ID [2], [3], [4], [5] and vehicle Re-ID [6].

Manuscript received 29 August 2023; revised 30 November 2023 and 22 January 2024; accepted 26 March 2024. Date of publication 17 May 2024; date of current version 29 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0102001 and in part by the National Natural Science Foundation of China under Grant 62101092. The Associate Editor for this article was H. Huang. (Corresponding author: Pingping Zhang.)

Xuehu Liu is with the School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China (e-mail: liuxuehu@whut.edu.cn).

Pingping Zhang is with the School of Artificial Intelligence, Dalian University of Technology, Dalian 116024, China (e-mail: zhpp@dlut.edu.cn).

Chenyang Yu and Huchuan Lu are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: yuchenyang@mail.dlut.edu.cn; lhchuan@dlut.edu.cn).

Xuesheng Qian is with CAS Intelligcloud Ltd., Shanghai 201250, China (e-mail: xuesheng.qian@intelligcloud.ai).

Xiaoyun Yang is with Remark Holdings, SE1 9PD London, U.K. (e-mail: xyang@remarkholdings.com).

Digital Object Identifier 10.1109/TITS.2024.3386914

In particular, person Re-ID plays an important role in safe communities, intelligent surveillance and criminal investigations. However, it is still a very challenging task, which is characterized by a wide variety of visual difficulties, such as viewpoint changes, complex backgrounds and object occlusions. Generally, person Re-ID can be categorized into image-based person Re-ID and video-based person Re-ID according to the types of input data. Nowadays, video-based person Re-ID has attracted increasing interest from researchers since the widespread deployment of video surveillance. Different from image-based person Re-ID, video-based person Re-ID takes video sequences as inputs rather than static images, which contain additional motion cues, pose variations and multi-view observations. Despite the advantages in terms of enriched visual representations, these information may bring some issues, such as more spatial noises and temporal misalignments. Therefore, how to fully exploit the abundant spatial-temporal information in sequences is very worthy of research in video-based person Re-ID.

Currently, researchers have explored various methods of processing video data to extract discriminative visual cues. Their efforts generally are based on three views of videos: spatial view, temporal view and spatial-temporal view. The intuitive comparison of different paradigms is shown in Fig. 1. Specifically, spatial view-based methods [9], [10], [11] leverage specialized 2D Convolutional Neural Networks (CNNs) to extract discriminative appearance features from each frame. Temporal view-based methods [12], [13], [14] design recurrent networks to mine dynamic information of videos via the inter-frame relationships. Meanwhile, spatial-temporal view-based methods [15], [16], [17], [18] exploit 3D CNNs to jointly explore spatial-temporal cues in a co-learning manner.

Authorized licensed use limited to: IEEE Xplore. Downloaded on August 20, 2024 at 13:03:37 UTC from IEEE Xplore. Restrictions apply.
However, those single-view methods lack multi-view observations and limit the full presentation of video information. Considering this, several approaches [19], [20], [21], [22], [23] employ dual-path networks that go beyond single-view observations by combining spatial and temporal learning. For example, Yang et al. [20] construct two Graph Convolutional Networks (GCN) to separately encode spatial and temporal information. Pan et al. [23] implement a two-branch architecture to separately learn the spatial appearance feature and temporal pose feature. Intuitively, those methods with different perspectives from the same sequence could yield more comprehensive video representations. In our work, we attempt to simultaneously capture three kinds of observations with view projectors from spatial, temporal and spatial-temporal views for a higher Re-ID accuracy. The paradigm of our method is shown in Fig. 1 (d).

In this paper, we propose a novel learning framework named TrigeMinal Transformer (TMT) for video-based person Re-ID. It aims at extracting robust video representations from spatial, temporal and spatial-temporal views. In each view, the inter-local relationships are utilized to assign different weights to local features for information enhancement. Meanwhile, the cross-view interactions are taken into consideration to aggregate multi-view cues and obtain final video representations. Specifically, our framework mainly consists of three key modules, as shown in Fig. 1 (d). First, a trigeminal feature extractor with View-wise Projectors (VP) is designed to adaptively transform raw videos from spatial, temporal, and spatial-temporal views. Second, three Self-view Transformers (ST), including spatial Transformer, temporal Transformer and spatial-temporal Transformer, are introduced to respectively enhance the view-wise features. Each ST exploits the inter-local relationships in the feature space and strengthens view-wise representations. Finally, a Cross-view Transformer (CT) is proposed to estimate the interactions among different views and aggregate them for the final video representation. Based on the above modules, our TMT can not only capture different characteristics in diverse views, but also fully aggregate multi-view information to generate more comprehensive video representations. Extensive experiments on four public benchmarks are conducted to demonstrate that our approach performs better than most state-of-the-art methods.

The main contributions of our work are as follows:

- We propose a novel learning framework named TrigeMinal Transformer (TMT) for video-based person Re-ID.
- We design a trigeminal feature extractor with view-wise projectors to transform raw videos into spatial, temporal, and spatial-temporal views for different observations.
- We introduce diverse Transformers into video-based person Re-ID. Three self-view Transformers are designed to enhance single-view features and a cross-view Transformer is designed to aggregate multi-view features.
- Experimental results on four public benchmarks demonstrate that our framework synthetically attains a better performance than most state-of-the-art methods.

The rest of this paper is organized as follows: Section II describes the related works in person re-identification. Section III presents the proposed TMT in detail. Section IV reports the results and analysis of extensive experiments, followed by the conclusion and future work in Section V.

II. RELATED WORK

Over the past decade, person Re-ID has achieved great success with the rise of deep learning. Nowadays, video-based person Re-ID has gained significant improvements and Transformer has been widely deployed in person Re-ID. Here, we give a brief review of these methods.

A. Video-Based Person Re-Identification

In video-based person Re-ID, several approaches have been proposed to extract robust video representations. Compared with static images, videos contain more information that can be observed from a spatial view, temporal view and spatial-temporal view. Therefore, some existing works [24], [25], [26] concentrate on extracting attentive spatial features with specialized extractors. Some approaches [27], [28], [29], [30] attempt to encode temporal information across multiple frames. In addition, some approaches [31], [32], [33], [34] attempt to obtain robust video representations in a co-learning manner. More specifically, in spatial view, Hou et al. [24] utilize a spatial generation network to complement the occluded appearance of pedestrians. Wu et al. [25] construct a structure-aware adjacency graph, exploiting the pose and the feature connections to refine regional features. Zhang et al. [26] extract multi-granularity spatial features guided by the global features. In temporal view, Chen et al. [28] disentangle videos into temporal coherent features and dynamic features by adversarial learning. Liu et al. [29] adaptively enhance and accumulate disentangled features with temporal reciprocal learning. Gu et al. [30] extract pedestrian motion information based on the position and appearance changes. In spatial-temporal view, Fu et al. [32] propose a spatial-temporal attention module for weighted feature aggregations. Yan et al. [33] construct hypergraphs with multiple granularities to capture spatial-temporal dependencies. Wang et al. [34] aggregate the multi-scale and multi-term spatial-temporal features utilizing a pyramid network. However, the perception focusing on a single view is limited. The appearance features from the spatial view and motion cues from the temporal view are conducive to complementary representations. Beyond the above single-view learning, we propose a view-based trigeminal framework to transform raw videos into spatial, temporal and spatial-temporal feature space for different observations. Meanwhile, we utilize self-view Transformers to fully explore discriminative cues.

B. Transformers for Person Re-Identification

Transformer [35] is initially designed to handle the sequential data for Natural Language Processing (NLP). Recently, Transformers have been brought into numerous computer vision tasks, which can capture the relational dependencies among contextual information and obtain integral representations with the global perception. For example, Dosovitskiy et al. [36] propose the Vision Transformer (ViT),
which applies self-attention operations among image patches and achieves promising results in image classification. Based on this, He et al. [37] introduce a jigsaw patch module into ViT for object Re-ID. Further, Zhang et al. [38] design a hierarchical aggregation Transformer to aggregate multi-scale features for image-based person Re-ID. Meanwhile, in video-based person Re-ID, He et al. [39] propose a dense interaction Transformer to model spatial-temporal information. Besides, Wu et al. [40] propose a contextual alignment vision Transformer to capture spatial-temporal clues. Tang et al. [41] propose a multi-stage spatial-temporal aggregation Transformer to fuse local and global features. Liu et al. [42] propose to deeply couple convolutional neural networks and Transformers towards spatial–temporal complementary learning. Compared with existing Transformers, our approach adaptively unifies spatial, temporal and spatial-temporal features with trigeminal view-wise projectors for video-based person Re-ID. In addition, we propose a novel cross-view Transformer that aggregates diverse observations from three views to obtain comprehensive video representations.

III. PROPOSED METHOD

Generally, spatial-temporal representation learning is treated as the key to tackling video-based person Re-ID. Our work explicitly explores the multi-view nature of videos to extract more discriminative representations. To this end, we present a trigeminal structure with spatial, temporal, and spatial-temporal views. As shown in Fig. 2, the framework mainly includes three key modules: trigeminal feature extractor, Self-view Transformer (ST), and Cross-view Transformer (CT). More specifically, a trigeminal feature extractor with View-wise Projectors (VP) transforms raw videos to generate diverse features. Afterward, ST explores the inter-local relations of view-wise features to enhance the spatial, temporal and spatial-temporal representations. Furthermore, CT aggregates multi-view observations to obtain the final video representation. We elaborate on these components in the following subsections.

A. Trigeminal Feature Extractor

Given a video sequence, the Restricted Random Sampling (RRS) [31] is adopted to generate T frames \(\{I_1, I_2, \ldots, I_T\}\) as the inputs of our framework. As shown in the left part of Fig. 2, our trigeminal feature extractor utilizes the ResNet-50 [43] to extract individual feature maps. We deploy the convolutional residual blocks from Res0_x to Res4_x in ResNet-50 as our basic network. After that, similar to the previous multi-branch work [44], three non-shared \(1 \times 1\) convolutional layers are utilized as view-wise embedding heads to generate three channel-reduced video feature tensors \((X^1, X^2, X^3)\). Each feature tensor \(X^i \in \mathbb{R}^{T \times H \times W \times C}\) \((i = 1, 2, 3)\) is with dimensions of \(T \times H \times W \times C\). Here, \(H\) and \(W\) are the height and width of the feature tensor. \(C\) and \(T\) are the reduced channel size and sequence length, respectively. Furthermore, we propose the view-wise projectors for diverse view representations.

View-Wise Projector: As shown in Fig. 3, our view-wise projector transforms video features into different views. It includes a view transformation and a view encoder.

Without loss of generality, in the view transformation, we first apply a Temporal Average Pooling (TAP) to the input video sequence \(I\), where \(G^s \in \mathbb{R}^{H \times W \times C}\), \(k \in [1, T]\), \(i \in [1, H]\) and \(j \in [1, W]\). Similarly, we apply a Spatial Average Pooling (SAP) to the feature tensor \(X^1\) to generate \(G^t\),

\[
G^t_{i,j} = \frac{\sum_{k=1}^{T} X^1_{k,i,j}}{T}, \tag{1}
\]

where \(G^s \in \mathbb{R}^{H \times W \times C}\), \(k \in [1, T]\), \(i \in [1, H]\) and \(j \in [1, W]\). Similarly, we apply a Spatial Average Pooling (SAP) to the feature tensor \(X^1\) to generate \(G^t\),

\[
G^t_k = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} X^1_{k,i,j}}{H \times W}, \tag{2}
\]

where \(G^t \in \mathbb{R}^{T \times C}\). Furthermore, we apply a Spatial-Temporal Average Pooling (STAP) to the feature tensor \(X^3\) to generate \(G^{st}\),

\[
G^{st}_{k,i,j} = \frac{\sum_{l_1=1}^{r_1} \sum_{l_2=1}^{r_2} X^3_{k,i,j}}{r_1 \times r_2 \times r_3}, \tag{3}
\]
where $G^H \in \mathbb{R}^{T \times H' \times W' \times C}$, $r_1 = \frac{T}{h}$, $r_2 = \frac{H}{H'}$ and $r_3 = \frac{W}{W'}$ are three positive integers larger than 1.

Apparently, the average pooling is hard to reduce the influence of noises in spatial, temporal and spatial-temporal, while resulting in poor representations. Therefore, we introduce attention mechanisms to assign weights for local features and generate discriminative observations by removing the noises in spatial, temporal and spatial-temporal. Firstly, each local feature of $X^1$ is transformed by a linear projection. We can formulate this by

$$\tilde{X}^s_{i,j} = \phi(X^s_{i,j}),$$ (4)

where $\phi$ represents the linear projection. After that, $G^s$ is utilized as the reference to estimate the attention weights $A^s_{i,j}$ for local features in temporal. Meanwhile, matrix multiplication and residual connection are conducted to obtain the attentive spatial feature $G^s \in \mathbb{R}^{H \times W \times C}$,

$$A^s_{i,j} = \frac{\exp(\cos(G^s_{i,j}, \tilde{X}^s_{i,j}))}{\sum_{i=1}^{T} \exp(\cos(G^s_{i,j}, \tilde{X}^s_{i,j}))},$$ (5)

$$\tilde{G}^s_{i,j} = \sum_{i=1}^{T} \tilde{X}^s_{i,j} A^s_{i,j} + G^s_{i,j}.$$ (6)

Noted that the attention structures in spatial-temporal and temporal views have similar structures, generating attentive features $G^st \in \mathbb{R}^{T \times H' \times W' \times C}$ and $G^t \in \mathbb{R}^{T \times C}$, respectively.

For view encoders, we employ a 2D Convolutional layer (2D-Conv), a Long Short-Term Memory (LSTM), and a 3D Convolutional layer (3D-Conv) to separately encode diverse view-wise features,

$$F^s = \Theta(G^s), \quad F^t = \Omega(G^t), \quad F^{st} = \Upsilon(G^{st}),$$ (7)

where $\Theta, \Omega, \Upsilon$ present the 2D-Conv, LSTM and 3D-Conv, respectively. In this way, our view-wise projector utilizes different structures to encode view-specific information. Specifically, we utilize a 2D convolutional layer to learn spatial local information, a LSTM to capture temporal long-term information, and a 3D convolutional layer to extract spatial-temporal continuous clues. To this end, we adaptively capture the discriminative spatial, temporal, and spatial-temporal representations after three view-wise projectors. In this way, our trigeminal feature extractor utilizes a partially shared network combined non-shared view-wise projectors to transform raw videos into multi-view representations, including spatial-view feature $F^s \in \mathbb{R}^{H \times W \times C}$, temporal-view feature $F^t \in \mathbb{R}^{T \times C}$ and spatial-temporal view feature $F^{st} \in \mathbb{R}^{T \times H' \times W' \times C}$.

**B. Self-View Transformer for Feature Enhancement**

Inspired by the strong capacity of Transformers, we introduce diverse Transformers to enhance the view-wise features. It captures the relationships of contextual information in view spaces for feature enhancement. The Self-view Transformers (ST) are multi-layer architectures, and each layer of the ST is composed of a Multi-Head Self-Attention (MHSA), a Feed-Forward Network (FFN), normalization and residual connections [35]. More specifically, taking the temporal enhancement as an example (shown in the left part of Fig. 4), ST can be formulated as follows. First, the feature $F^t \in \mathbb{R}^{T \times C}$ from the temporal-view projector is added to its position information $r$ [36], which is a learnable embedding matrix, then is passed through the MHSA. In the $h$-th head of MHSA, the feature is fed into three linear transformations to generate features $Q, K$ and $V$. Here, $Q, K, V \in \mathbb{R}^{T \times d}, d = \frac{C}{N_h}$ and $N_h$ is the total number of heads. Utilizing them, the self-attention operation in MHSA is defined as:

$$M^t_h = \sigma(QK^T/\sqrt{d})V,$$ (8)

where $\sigma$ is the softmax activation function and $(\cdot)^T$ means the transposition operation. When assisted by effective self-attention, temporal representations are enhanced by capturing the relationships among multiple frames. The outputs of multiple heads ($M^t_1, \cdots, M^t_{N_h}$) are concatenated to be $M^t \in \mathbb{R}^{T \times C}$ as the global interaction feature. Afterward, $M^t$ is followed by a frame-wise FFN with residual connections and a normalization operation,

$$S' = \Phi(F^t + M^t).$$ (9)

where $\Phi$ is the frame-wise FFN, which encodes the frame-wise features to strengthen the temporal representations.
In the same way, the appearance representations from the spatial view are enhanced with the relational dependencies among contextual information. The integrative representations with the global perception from the spatial-temporal view are enhanced by capturing the relationships across regions and frames. In our work, we deploy three STs to model the relationships in spatial, temporal, and spatial-temporal views for feature enhancement, respectively. The relationships among local features are beneficial to better representations.

C. Cross-View Transformer for Feature Aggregation

With the aforementioned ST, the appearance, motion and global representations have been enhanced in spatial, temporal and spatial-temporal views. Furthermore, three kinds of view features are aggregated to obtain comprehensive video representations. Actually, the interactions across multiple views are incredibly valuable. Considering this fact, we propose a CT to model the interactions among spatial, temporal and spatial-temporal information for feature aggregation.

The right part of Fig. 4 shows the detailed structure of our proposed CT. Two Multi-Head Cross Attention (MHCA) layers are designed to estimate the interactions between view features, and a Feed-Forward Network (FFN) is utilized to encode aggregated information for the final video representation. Specifically, the spatial-temporal feature $S^{st} \in \mathbb{R}^{T \times H \times W \times C}$ is pooled by the Global Average Pooling (GAP) to generate the coarse video representation $\hat{Z} \in \mathbb{R}^{C}$. Based on this, in one MHCA, we combine $\hat{Z}$ and $S^{t} \in \mathbb{R}^{H \times W \times C}$ to generate the cross spatial attention. Meanwhile, $\hat{Z}$ and $S^{t} \in \mathbb{R}^{T \times C}$ are learned by another MHCA. Formally, in the $h$-th head of MHCA, six linear projections are applied to generate $Q^{st}_s, Q^{st}_t, K_s, V_s, V_t$, where $Q^{st}_s, Q^{st}_t \in \mathbb{R}^{1 \times d}$ are from $\hat{Z}$.

$$K_s, V_s \in \mathbb{R}^{H \times W \times d},$$

After that, the interactions between different view-wise features can be estimated as:

$$A^s_h = \sigma \left( \frac{Q^s_h K^s}{\sqrt{d}} \right), \quad A^t_h = \sigma \left( \frac{Q^t_h K^t}{\sqrt{d}} \right),$$

where $A^s_h \in \mathbb{R}^{1 \times HW}$ and $A^t_h \in \mathbb{R}^{1 \times T}$. By this, the interactions from spatial and temporal views are explored attentively with the guidance of spatial-temporal features. Furthermore, in each head of MHCA, the attentions $A^s_h$ and $A^t_h$ are assigned to corresponding local-wise features,

$$\hat{Z}^s = A^s_h \otimes V_s, \quad \hat{Z}^t = A^t_h \otimes V_t.$$  

Furthermore, the concatenation of multi-head outputs $(\hat{Z}^s_1, \cdots, \hat{Z}^s_N, \hat{Z}^t_1, \cdots, \hat{Z}^t_N)$ and $(\hat{Z}^s_1, \cdots, \hat{Z}^s_N, \hat{Z}^t_1)$ can be represented as $\hat{Z}^s$ and $\hat{Z}^t$, which have the same size of $\hat{Z}$.

After that, with a FFN, the multi-view features are aggregated to generate the final video representation $Z \in \mathbb{R}^{C}$,

$$Z = \Psi(\hat{Z} + \hat{Z}^s + \hat{Z}^t).$$

where $\Psi$ is the FFN, which consists of two fully-connected layers with GELU activations [45] and a residual connection. Based on the above operations, the diverse view-wise features are aggregated to generate more comprehensive representations for video-based person Re-ID.

D. Multi-Stage Supervision

As shown in Fig. 2, our supervision is deployed at four stages: after backbone, after VP, after ST and after CT. At each stage, we combine the cross-entropy loss [46] and triplet
loss [47] for network optimization. Formally, the cross-entropy loss can be computed by

\[ L_{cls} = \frac{1}{N} \sum_{n=1}^{N} y_i \log \frac{\exp(W_i X_n)}{\sum_j \exp(W_j X_n)}, \tag{13} \]

where \( X_n \) is the feature vector of the \( n \)-th sequence in the training set and \( y_i \) is its corresponding ground-truth label. \( N \) means the number of sequences and \( J \) denotes the total number of classes. In the experiments, the cross-entropy loss with label smoothing [46] can learn the identify-specific representation and avoid overfitting. Meanwhile, we utilize the triplet loss with batch hard mining [47] that draws the positives and pushes the negatives in the feature spaces to improve the ranking performance. For this, we construct the feature set \( \Omega = \{ F_a, F_p, F_n \} \) in a mini-batch. Here, \( F_a, F_p \) and \( F_n \) denote the anchor feature, positive feature, and negative feature respectively. Thus, the triplet loss can be defined as

\[ L_{tri} = \frac{1}{|\Omega|} \sum_{F_a, F_p, F_n \in \Omega} [\epsilon + d(F_a, F_p) - d(F_a, F_n)], \tag{14} \]

where \( \{x\}^+ = \max(x, 0) \) is the hinge loss, \( \epsilon \) is a distance margin and is set to 0.3 in our experiments. \( d(\cdot, \cdot) \) represents the cosine distance between features. Finally, the whole loss function can be formulated as:

\[ L_{total} = \frac{1}{4} \sum_{m=1}^{4} (L_{cls}^m + L_{tri}^m), \tag{15} \]

where \( m \) is the ordinal number of our multi-stage supervision.

IV. EXPERIMENTS

A. Datasets and Evaluation Protocols

In this paper, we adopt four widely-used person Re-ID benchmarks to evaluate our proposed method, i.e., iLIDS-VID [48], MARS [49], DukeMTMC-VID [50] and LS-VID [27]. iLIDS-VID is a small dataset with two cameras capturing images. iLIDS-VID has 600 video sequences of 300 different identities. MARS is one of the large-scale datasets and consists of 1,261 identities around 18,000 video sequences. Note that, there are around 3,200 distracted sequences in the dataset to simulate actual conditions. DukeMTMC-VID is another large-scale dataset. It comprises 4,832 sequences from 1,812 identities including 408 distractor identities. LS-VID is a recently released large-scale dataset with 4,832 sequences collected from 1,812 identities. For evaluation, we follow previous works and adopt the cumulative matching characteristic table (Rank-1 and Rank-5 termed as R-1 and R-5) and mean Average Precision (mAP). Note that mAP is not reported on iLIDS-VID dataset in TABLE I because each query only has one right match in galleries.

B. Implementation Details

In this work, we conduct experiments with the PyTorch toolbox. The experimental devices include an Intel i4790 CPU and two NVIDIA RTX3090 GPUs (24G memory). Each image in a video sequence is resized to 256 × 128 and augmented by random cropping, horizontal flipping and random erasing. The ResNet-50 [43] pre-trained on the ImageNet classification dataset [51] is used as our backbone network. Following previous work [52], we remove the last spatial down-sampling operation to increase the feature resolution. Experimentally, we set the \textit{batchsize} = 32, the length of video sequence \( T = 8 \), and the size of reduced channel \( C = 512 \). In a mini-batch, we randomly select 8 person identities where each identify samples 4 video sequences. The whole network is updated by the Adam [53] algorithm with an initial learning rate of 0.0003, a weight decay \( 5 \times 10^{-4} \). During training, the learning rate is decayed by 10 at every 70 epochs until 500 epochs.

C. Compared With State-of-the-Arts

In this section, our proposed TMT is compared with other state-of-the-art methods. The results are reported in TABLE I and TABLE II. Existing approaches are coarsely separated into four groups, i.e., works in spatial view, spatial-temporal view and multi-view. From the results, one can observe that, on MARS and iLIDS-VID, our proposed TMT attains more expressive performances than other methods. In addition, our method achieves the highest mAP accuracy on DukeMTMC-VID and LS-VID, and achieves the highest R-1 accuracy on MARS and DukeMTMC-VID. Furthermore, one can find the following facts:

1. The spatial view-based methods [9], [10], [11], [25], [26] concentrate on designing effective spatial feature extractors, such as attribute predictor [10], pose estimator [11] and graph modeling [25]. In contrast, our proposed TMT considers the transformation of spatial view with a view-wise projector, reducing the noise of backgrounds. It can enhance spatial-view features and achieve better performances.

2. Our method outperforms GRL by 1.3% mAP on MARS. Compared with existing temporal learning methods [14], [27], [29], [54], the significant improvements can be attributed to the use of Transformers that capture the inter-frame relationships and encode better temporal representations.

3. Most existing spatial-temporal view-based methods [17], [18], [32], [33] lack the multi-view observations and limit the presentation ability of video information. For instance, the outstanding method MGH [33] extracts the multi-granularity spatial-temporal information with hyper-graphs. It achieves 85.6% and 90.0% R-1 accuracies on iLIDS-VID and MARS, respectively. In contrast, our method combines three kinds of view representations for more comprehensive video observations. It helps our framework to gain 5.7% and 1.8% R-1 improvements on iLIDS-VID and MARS than MGH.

4. DIL [39] also employs the spatial-temporal Transformer to strengthen the features after ResNet-50. Different from it, our proposed trigeminal structure can encode and aggregate multi-view features, and outperforms DIL with 1.1% improvements in R-1 accuracy on MARS. We note that CAViT utilizes multi-scale patch learning and MSTAT deploys multi-stage aggregation Transformers. They are excellent works and achieve impressive performances in video-based person ReID. Different from them, our TMT utilizes view-wise projectors to extract different view-specific features, following self-view Transformers for enhancement and a cross-view Transformer for aggregation. As shown in TABLE I and TABLE II, our TMT outperforms CAViT by 1.0% and 1.8% mAP on MARS and LS-VID datasets and surpasses MSTAT by 1.2% and 1.1% mAP.
TABLE I
PERFORMANCE (%) COMPARISON ON iLIDS-VID [48], MARS [49] AND DUKEMTMC-VID [50]. COMPARED METHODS ARE SEPARATED INTO FOUR GROUPS, i.e., SPATIAL VIEW (S), TEMPORAL VIEW (T), SPATIAL-TEMPORAL (ST) VIEW AND MULTIPLE VIEWS (MV). THE TEXTS IN BOLD AND UNDERLINED HIGHLIGHT THE BEST AND SECOND PERFORMANCES, RESPECTIVELY.

| Methods          | iLIDS-VID | MARS    | DukeMTMC-VID |
|------------------|-----------|---------|--------------|
|                  | R-1 | R-5 | mAP | R-1 | mAP |
| S                |     |     |     |     |     |
| COSAM [9]        | 77.8 | 97.3 | 79.9 | 84.9 | 94.1 | 95.4 |
| AGRL [25]        | 83.7 | 95.4 | 81.1 | 89.8 | 94.2 | 96.7 |
| Attribute [10]   | 86.3 | 97.4 | 78.2 | 87.0 | -    | -    |
| MG-RAFA [26]     | 88.6 | 98.0 | 85.9 | 88.8 | -    | -    |
| CTL [11]         | 89.7 | 97.0 | 86.7 | 91.4 | -    | -    |
| T                |     |     |     |     |     |
| STMP [14]        | 84.3 | 96.8 | 72.7 | 84.4 | -    | -    |
| AMOC [54]        | 68.7 | 94.3 | 52.9 | 68.3 | -    | -    |
| GLTR [27]        | 86.0 | -   | 78.5 | 87.0 | 93.7 | 96.3 |
| GRL [29]         | 90.4 | 98.3 | 84.8 | 91.0 | -    | -    |
| CAVIT [40]       | 93.3 | -   | 87.2 | 90.8 | -    | -    |
| ST               |     |     |     |     |     |
| PersonVLAD [15]  | 70.7 | 88.2 | 64.7 | 82.8 | -    | -    |
| STA [32]         | -    | -   | 80.8 | 86.3 | 94.9 | 96.2 |
| MGH [33]         | 85.6 | 97.1 | 85.8 | 90.0 | -    | -    |
| SSN3D [18]       | -    | -   | 86.2 | 90.1 | 96.3 | 96.8 |
| AP3D [17]        | 88.7 | -   | 85.6 | 90.7 | 96.1 | 97.2 |
| DIL [39]         | 92.0 | 98.0 | 87.0 | 90.8 | 97.1 | 97.6 |
| MSTAT [41]       | 93.3 | 99.3 | 85.3 | 91.8 | 96.4 | 97.4 |
| ST               |     |     |     |     |     |
| M3D [19]         | 86.7 | 98.0 | 85.5 | 88.9 | 93.7 | 95.5 |
| STGCN [20]       | -    | -   | 83.7 | 90.0 | 95.7 | 97.3 |
| STRF [21]        | 89.3 | -   | 86.1 | 90.3 | 96.4 | 97.4 |
| STMN [22]        | -    | -   | 84.5 | 90.5 | 95.9 | 97.0 |
| RGCCN [23]       | 90.2 | 98.5 | 86.5 | 91.1 | 96.5 | 97.1 |
| MV               |     |     |     |     |     |
| TMT (Ours)       | 91.3 | 98.6 | 86.5 | 91.8 | 97.5 | 97.8 |

TABLE II
PERFORMANCE (%) COMPARISON ON LS-VID [27]

| Methods | LS-VID | mAP | R-1 |
|---------|--------|-----|-----|
| S       |        |     |     |
| SINet [55] | 79.6 | 87.4 |
| T       |        |     |     |
| GLTR [27] | 44.3 | 63.1 |
| TCLNet [56] | 70.3 | 81.5 |
| CAVIT [40] | 79.2 | **89.2** |
| ST      |        |     |     |
| AP3D [17] | 73.2 | 84.5 |
| BiCNet [57] | 75.1 | 84.6 |
| MV      |        |     |     |
| STMN [22] | 69.2 | 82.1 |
| TMT (Ours) | **81.0** | **88.9** |

mAP on MARS and DukeMTMC-VID datasets. The performance improvement demonstrates that incorporating diverse information from spatial, temporal, and spatial-temporal views is crucial for video-based re-identification.

(5) Some methods [19], [20], [21], [22], [23] separate spatial and temporal learning to extract diverse visual representations. For example, STMN [22] first leverages the spatial and temporal memories, and then refines different patterns in person videos. STRF [21] factorizes the video feature to capture the appearance and motion cues. RGCN [23] implements two branches to separately extract the spatial and temporal pose features. However, these methods aggregate diverse representations with direct concatenation or summation. In our work, we estimate the interactions across multiple views to yield more comprehensive video representations. Thus, our proposed method surpasses RGCN by 1.0% and 0.7% in terms of mAP and R-1 accuracy on DukeMTMC-VID, respectively.

In summary, our method represents a video from different views and generates more comprehensive and robust representations. The higher performances validate the superiority of our proposed method, indicating the ability of our TMT to obtain more comprehensive video representations.

D. Ablation Study
To investigate the effectiveness of our proposed modules, we conduct ablation experiments on two public large-scale benchmarks: MARS and DukeMTMC-VID.

Effectiveness of Key Modules. A comprehensive ablation study is performed to analyze the effects of three proposed modules: VP, ST, and CT. Results are shown in TABLE III. The first row shows the results of the baseline method which adopts ResNet-50 with three embedding heads. The features from three heads followed by the GAP and Temporal Average Pooling (TAP) are concatenated as the final video representation. As shown in the 2nd, 3rd and 4th rows of TABLE III, when we gradually add the proposed key modules to the baseline, the resulting performances are improved. Specifically, compared with the simple embedding heads, our proposed VP helps to reduce the influence of noises from backgrounds and disrupted frames. Thus, more discriminative representations after VPs improve the mAP by 0.9% and 0.5% than the baseline on MARS and DukeMTMC-VID, respectively. In addition, ST enables view-wise representations to better...
capture the contextual relationships and temporal dependencies for feature enhancement, resulting in better performances than the baseline method. They gain 1.1% and 2.0% improvements in terms of mAP and R-1 accuracy on MARS, and 0.7% and 1.0% improvements on DukeMTMC-VID. Meanwhile, we adopt a CT to replace the concatenation to aggregate multiple view features, which helps to achieve 1.3% and 1.0% improvements in mAP on MARS and DukeMTMC-VID, respectively. These results indicate that the interactions among different views help obtain a more comprehensive video representation. Furthermore, ablation results with different combinations of our proposed modules are shown in the 5th, 6th, and 7th rows of TABLE III. The combination of the three modules results in further performance improvements. Particularly, as shown in the last row of TABLE III, our method with all the key modules can improve the baseline by 3.0% and 1.9% R-1 accuracy on MARS and DukeMTMC-VID, respectively. These results demonstrate the effectiveness of our key modules.

Effect of View-Wise Representations: To verify the effects of view-wise representations, experiments are conducted on three independent views with our VP, ST and CT. The results are shown in TABLE IV. The baseline method is shown in the 1st row of TABLE IV. It utilizes three non-shared embedding heads for three different feature representations. One can see that the performances are almost the same, and the aggregation by feature concatenation has little improvement. However, in our framework, three view-wise projectors are utilized to extract diverse view features. The performance of resulting models with three views can be promoted to different extents, better than the baseline method. Meanwhile, further improvements can be achieved when ST is deployed after VP for feature enhancement. Besides, the aggregation of three views shows higher results than single views, which confirms our statement: A video is worth three views. In addition, we apply CT to replace the concatenation for feature aggregation and finally attain the best performance on MARS. Compared with the concatenation after VP and ST, our CT models the interactions among diverse views and helps to gain 0.5% mAP and 0.4% R-1 accuracy on MARS. These incremental performances further indicate that our CT is beneficial to aggregate multi-view features for robust video representations.

**Effect of Spatial Feature Sizes in Spatial ST:** Results in Fig. 5 (a) present the effect of different spatial feature sizes in spatial ST on MARS. Experimentally, we vary the strides of convolutional kernels in the 2D-Conv of VP. The performance improves as the spatial feature sizes increase. For the best performance, we utilize the $16 \times 8$ spatial size as our default settings in our experiments.

**Effect of Video Sequence Lengths in Temporal ST:** Results in Fig. 5 (b) present the effect of different video sequence lengths on MARS. To achieve this goal, we change the video sequence lengths from 4 to 10. From the results, one can see that the temporal ST is sensitive to the length of video sequences. While long sequences bring more temporal information, they also bring more noise. When varying the length of video sequences to 8, the temporal ST gets the best performance in terms of both mAP and R-1 accuracy.

**Effect of Different Depths in Transformers:** The depth of Transformers is an important factor for exploring the relationships of local features. In Fig. 6, we present the performances...
E. Visualization Analysis

Visualization of Attention Weights in VP: We show the attention weights obtained from Equ. 5 in view-wise projectors. In spatial VP, the attention $A^s$ is utilized to aggregate the local information for discriminative spatial view representations. As shown in Fig. 7 (b), compared with the red boxes in each image of a video, our spatial VP pays more attention to the pedestrian body rather than useless contexts. It gathers meaningful information across frames into spatial features. In temporal VP, spatial attention can suppress the noises from backgrounds by focusing on the foreground, as shown in Fig. 7 (c). Therefore, better feature vectors can be generated from a temporal view with our VP. Furthermore, the spatial-temporal VP captures more cues relevant to identification from spatial-temporal features. As observed from Fig. 7 (d), the spatial-temporal attention highlights the local discriminative regions to keep valuable information, which appears in the brown boxes. These visualizations suggest that

with different depths of Transformers on MARS. We add the spatial/temporal/spatial-temporal ST and CT to the baseline method respectively. From the results, one can see that the R-1 accuracy is almost all the best for ST and CT when the depth is 3. To trade off the computation and performance, in this work, we set all the depth of Transformers to 3.

Effects of the Multi-Stage Supervision: Here, we verify the effectiveness of our multi-stage supervision. More specifically, there are four supervised levels, i.e., the backbone supervision, VP supervision, ST supervision and CT supervision. The results are presented in TABLE V. The use of four-stage supervision has the best performance on MARS and DukeMTMC-VID. The reason may be that it provides more supervision information, thus better optimizing the whole framework. Meanwhile, we find that the performances will be severely degraded on both datasets when only employing CT supervision.

TABLE V

| Different Supervised Stages | MARS mAP | R-1 | DukeMTMC-VID mAP | R-1 |
|-----------------------------|----------|-----|------------------|-----|
| CT ST VP Backbone           |          |     |                  |     |
| ✓ ✓ ✓ ✓                      | 86.5     | 91.8| 97.5             | 97.8|
| ✓ ✓ ✓ ×                      | 85.5     | 90.5| 95.0             | 94.9|
| ✓ ✓ × ×                      | 84.7     | 89.8| 93.9             | 93.6|
| ✓ × × ×                      | 50.4     | 65.7| 67.6             | 68.2|

Visualization of Retrieval Results: The retrieval results for a hard query sample by utilizing different methods are shown in Fig. 8. As can be observed, it is difficult for the baseline method to re-identify the given person due to the minimal inter-class difference. In our approach, we employ three view-wise Transformers to enhance the view representations by capturing the relationships. As shown in the 2nd, 3rd, and 4th columns of Fig. 8, compared with the retrieval results of the baseline, the rank accuracies are improved in various degrees by our view-wise Transformers. More specifically, the temporal ST mines the dependencies among frames for better motion representations. The spatial ST enhances the appearance representations by exploring the relational dependencies among contextual information. The spatial-temporal ST captures the relationships across regions and frames for integral representations with global perceptions. Moreover, with the help of the interactions between multiple views, the aggregated features from spatial, temporal and spatial-temporal views by CT exhibit the powerful ability for video representations. Therefore, the rank accuracy is further boosted, as shown in the 5th column of Fig. 8, which indeed demonstrates that a video is worth three views.
V. CONCLUSION AND FUTURE WORKS

In this paper, we propose a novel learning framework named Trigeminal Transformer for video-based person Re-ID. We use a trigeminal feature extractor to capture spatial, temporal and spatial-temporal information, respectively. In addition, view-wise projectors are designed to extract view-specific features effectively. Meanwhile, Self-view Transformers (ST) are introduced to explore the relationships of local features for information enhancement. A Cross-view Transformer (CT) is proposed to model the interactions of multiple views and integrate the view-wise features for more comprehensive representations. Based on these modules, we present a video from spatial, temporal and spatial-temporal views, and verify that a video is worth three views for video-based person Re-ID. Experiments on four benchmarks demonstrate that our approach performs better than the most state of the arts. In the future, we will explore how to generalize our Trigeminal Transformer to other video-based vision tasks such as action recognition and video segmentation, etc. Besides, we attempt to construct a unified vision model with three-view learning following specific task predictors to address multiple video-based vision tasks.

REFERENCES

[1] M. Tu et al., “Multi-granularity mutual learning network for object re-identification,” IEEE Trans. Intell. Transp. Syst., vol. 23, no. 9, pp. 15178–15189, Sep. 2022.
[2] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, “A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 3, pp. 523–536, Mar. 2019.
[3] Q. Leng, M. Ye, and Q. Tian, “A survey of open-world person re-identification,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 4, pp. 1092–1106, Apr. 2020.
[4] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, “Deep learning for person re-identification: A survey and outlook,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 2872–2893, Jun. 2022.
[5] M. Xu, H. Guo, Y. Jia, Z. Dai, and J. Wang, “Pseudo label rectification with joint camera shift adaptation and outlier progressive recycling for unsupervised person re-identification,” IEEE Trans. Intell. Comput., vol. 24, no. 3, pp. 3395–3406, Mar. 2023.
[6] M. Li, J. Liu, C. Zheng, X. Huang, and Z. Zhang, “Exploiting multi-view part-wise correlation via an efficient transformer for vehicle re-identification,” IEEE Trans. Multimedia, vol. 25, pp. 919–929, 2023.
[7] Y. Chen, S. Zhang, F. Liu, C. Wu, K. Giao, and Z. Qi, “DVHN: A deep hashing framework for large-scale vehicle re-identification,” IEEE Trans. Intell. Transp. Syst., vol. 24, no. 9, pp. 9268–9280, Sep. 2023.
[8] Z. Yu, Z. Huang, J. Pei, L. Tahsin, and D. Sun, “Semantic-oriented feature coupling transformer for vehicle re-identification in intelligent transportation system,” IEEE Trans. Intell. Transp. Syst., vol. 25, no. 3, pp. 2803–2813, Mar. 2024.
[9] A. Subramaniam, A. Nambiar, and A. Mittal, “Co-segmentation inspired attention networks for video-based person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 562–572.
[10] 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 Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4913–4922.
[11] J. Liu, Z.-J. Zha, W. Wu, K. Zheng, and Q. Sun, “Spatial-temporal correlation and topology learning for person re-identification in videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 4368–4377.
[12] N. McLaughlin, J. M. Del Rincon, and P. Miller, “Recurrent convolutional network for video-based person re-identification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1325–1334.
[13] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, “Video person re-identification by temporal residual learning,” IEEE Trans. Image Process., vol. 28, no. 3, pp. 1366–1377, Mar. 2019.
[14] Y. Liu, Z. Yuan, W. Zhou, and H. Li, “Spatial and temporal mutual promotion for video-based person re-identification,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 8786–8793.
[15] L. Wu, Y. Wang, L. Shao, and M. Wang, “3-D PersonVLAD: Learning deep global representations for video-based person reidentification,” IEEE Trans. Neural Netw. Learn. Syst., vol. 30, no. 11, pp. 3347–3359, Nov. 2019.
[16] J. Liu, Z.-J. Zha, X. Chen, Z. Wang, and Y. Zhang, “Dense 3D-convolutional neural network for person reidentification in videos,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 15, no. 1s, pp. 1–19, Jan. 2019.
[17] X. Gu, H. Chang, B. Ma, H. Zhang, and X. Chen, “Appearance-preserving 3D convolution for video-based person re-identification,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 228–243.
[18] X. Jiang, Y. Qiao, and J. Yan, “SSN3D: Self-separated network to align parts for 3D convolution in video person re-identification,” in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 2, 2021, pp. 1691–1699.
[19] J. Li, S. Zhang, and T. Huang, “Multi-scale temporal cues learning for video person re-identification,” IEEE Trans. Image Process., vol. 29, pp. 4461–4473, 2020.
[20] J. Yang, W.-S. Zheng, Q. Yang, Y.-C. Chen, and Q. Tian, “Spatial-temporal graph convolutional network for video-based person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2020, pp. 3289–3299.
[21] 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 Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 152–162.
[22] C. Eom, G. Lee, J. Lee, and B. Ham, “Video-based person re-identification with spatial and temporal memory networks,” in Proc. IEEE Int. Conf. Comput. Vis., Jul. 2021, pp. 12036–12045.
[23] H. Pan et al., “Pose-aided video-based person re-identification via recurrent graph convolutional network,” IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 12, pp. 7183–7196, Dec. 2023.
[24] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, “VRSTC: Occlusion-free video person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 7183–7192.
[25] Y. Wu, O. E. F. Bourahla, X. Li, F. Wu, Q. Tian, and X. Zhou, “Adaptive graph representation learning for video person re-identification,” IEEE Trans. Image Process., vol. 29, pp. 8821–8830, 2020.
[26] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, “Multi-granularity reference-aided attentive feature aggregation for video-based person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10407–10416.
[27] J. Li, S. Zhang, J. Wang, W. Gao, and Q. Tian, “Global-local temporal representations for video person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3958–3967.
[28] 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 Proc. Eur. Conf. Comput. Vis., 2020, pp. 660–676.
[29] X. Liu, P. Zhang, C. Yu, H. Lu, and X. Yang, “Watching you: Global-guided reciprocal learning for video-based person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 13334–13343.
[30] X. Gu, H. Chang, B. Ma, and S. Shan, “Motion feature aggregation for video-based person re-identification,” IEEE Trans. Image Process., vol. 31, pp. 3908–3919, 2022.
[31] S. Li, S. Bak, P. Carr, and X. Wang, “Diversity regularized spatiotemporal attention for video-based person re-identification,” in Proc. AAAI Conf. Artif. Intell., 2019, pp. 8287–8294.
[32] Y. Yan et al., “Learning multi-granular hypergraphs for video-based person re-identification,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 2899–2908.
[33] Y. Wang, P. Zhang, S. Gao, X. Geng, H. Lu, and D. Wang, “Pyramid spatial-temporal aggregation for video-based person re-identification,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 12006–12015.
Pingping Zhang received the B.E. degree in mathematics and applied mathematics from Henan Normal University (HNU), Xinxiang, China, in 2012, and the Ph.D. degree in signal and information processing from Dalian University of Technology (DUT), Dalian, China, in 2020. He is currently an Associate Professor with the School of Artificial Intelligence, DUT. His research interests include deep learning, saliency detection, person re-identification, and semantic segmentation.

Chenyang Yu received the B.E. degree in measurement and control technology and instrument and the M.E. degree in detection technology and automatic equipment from Northeastern University (NEU), Shenyang, China, in 2013 and 2017, respectively. He is currently pursuing the Ph.D. degree in signal and information processing with Dalian University of Technology (DUT), Dalian, China. His research interests include deep learning and person re-identification.

Xuesheng Qian is currently a Senior Researcher with Fudan University Smart City Research Center, a Technical Consultant of China Science Intell-Cloud Technology Company Ltd., and a member of Macau Institute of Systems Engineering. He is mainly engaged in research on the application of new information technology in business innovation and society management. He is a Senior Member of China Simulation Federation and the Vice Director of the Neuro-Mentalism and Neuro-Engineering Seminar, Society of Management Science and Engineering of China.

Xiaoyun Yang received the Ph.D. degree in visual computing from Ulster University in 2006. He is currently the Director of Remark Holdings. His research interests include applying computer vision research to real-world applications, recently focusing on the area of object detection, tracking, recognition, and semantic segmentation.