Robust Video-based Person Re-Identification by Hierarchical Mining

Zhikang Wang, Lihuo He, Xiaoguang Tu, Jian Zhao, Xinbo Gao, Shengmei Shen, Jiashi Feng

***Abstract*—Video-based person re-identification (Re-ID) aims at retrieving the person through the video sequences across non-overlapping cameras. Some characteristics of pedestrians are not consecutive across frames due to the variations of viewpoints, postures, and occlusions over time. However, existing methods ignore such data peculiarity and the networks tend to only learn those salient consecutive characteristics among frames in video sequences. As a result, the learned representations fail to cover all the characteristics of pedestrians, thus lacking integrity and discrimination. To tackle this problem, we present a novel deep architecture termed Hierarchical Mining Network (HMN), which mines as many pedestrians' characteristics by referring to the temporal and intra-class knowledge. It consists of a novel Attentive Temporal Module (ATM) and a Dynamic Supervising Branch (DSB), with a Balancing Triplet Loss (BTL) assisting the training. The proposed ATM, with pedestrian perceiving capacity, is capable of evaluating each activation of features through temporal analysis, so that the temporally scattered characteristics of pedestrians can be better aggregated and the contaminated ones can be eliminated. Then, the DSB along with the BTL further enhances the integrity of representations by multiple supervision. Specifically, the DSB perceives the diversities of intra-class samples in each mini-batch and generates targeted supervising signals for them, in which process the BTL guarantees the signals with smaller intra-class variations and larger inter-class variations. Comprehensive experiments on two video-based datasets, *i.e.*, MARS, and DukeMTMC-VideoReID, demonstrate the contribution of each component and the superiority of the proposed HMN over the state-of-the-arts. Benchmarking our model on three popular image-based datasets, *i.e.*, Market1501, DukeMTMC-Reid, and MSMT17 additionally verifies the promising generalizability of the proposed DSB and BTL.**

***Index Terms*—Person Re-ID, Hierarchical Mining, Dynamic Supervising, Temporal Attention.**

I. Introduction

*Lihuo He and Xinbo Gao are the corresponding authors.

Zhikang Wang is with School of Electronic Engineering, Xidian University, Pensees Singapore Institute, Pensees Pte Ltd, and the National University of Singapore. E-mail: zkwang00@gmail.com.

Lihuo He is with the School of Electronic Engineering, Xidian University. E-mail: lihuo.he@gmail.com.

Xiaoguang Tu is with Aviation Engineering Institute at Civil Aviation Flight University of China.

Jian Zhao is with Institute of North Electronic Equipment, Beijing, China. Homepage: https://zhaoj9014.github.io/. E-mail: zhaojian90@u.nus.edu.

Xinbo Gao is with the School of Electronic Engineering, Xidian University. E-mail: E-mail: xbgao@mail.xidian.edu.cn.

Shengmei Shen is with with Pensees Singapore Institute, Pensees Pte Ltd.

Jiashi Feng is with Department of Electrical and Computer Engineering, National University of Singapore, Singapore.

Fig. 1. Examples sequences in challenging scenes. The consecutive and inconsecutive characteristics are framed out with green boxes and orange boxes, respectively. The consecutive ones are stable and consistent among frames. As for the inconsecutive ones, their continuities and dynamics are broken by the various distractors. They may randomly distribute among frames without much dynamic information.

PERSON re-identification (Re-ID) aims at retrieving the same pedestrian of a camera view from other non-overlapping ones in single images or video sequences. Due to its wide practical applications like surveillance [1], activity analysis [2], and tracking [3, 4], a lot of previous works [5, 6, 7, 8, 9, 10] have been proposed. Compared with single images, video sequences contain richer appearance information and additional temporal cues, often yielding higher model accuracy even in challenging cases. However, as shown in Fig. 1, various distractors break the continuities and dynamics of some characteristics, leading to inconsecutive and inconspicuous. Therefore, how to generate representations with not only discrimination but also such high characteristic integrity through the sequence is a major challenge to video-based person Re-ID. Previous methods try to achieve by the temporal analysis and metric learning.

Temporal cues that are unique to video sequences are crucial to representation learning in video-based person Re-ID. Most methods extract frame-level features firstly and then adopt various aggregation mechanisms, *e.g.*, temporal pooling, or Recurrent Neural Networks (RNNs) [11, 12], to incorporate temporal information in the final representations. For instance, [8, 13] feed the frame-level features into the RNNs for temporal aggregation. Although RNNs are capable of learning long-term dependencies, learning by simply referring to the temporal prior knowledge would lead to representations only

concentrating on those consecutive and salient characteristics among frames while losing some other potentially valuable ones. [14, 15] adopt *Average Pooling* and *Max Pooling* operations along the temporal dimension for aggregation. However, they either introduce too much noise or discard too many valuable characteristics. To improve the discrimination of the representations, the attention mechanism has been popularly applied to Re-ID tasks [16, 17, 18, 19]. Generally, these works generate weights for frame-level features by temporal analysis and then aggregate the weighted features through *Pooling* operation. As valuable characteristics and distracting factors are often distributed discretely among the frames, generating frame-level weights for aggregation would inevitably weaken the role of the valuable characteristics of those badly contaminated frames in the final whole representation, *i.e.* leading to poor representation integrity.

For improving the discrimination and integrity of representations, deep metric learning is also much adopted to strengthen person Re-ID networks with various loss functions. Triplet loss [20] aims to pull the intra-class distances closer and simultaneously push the inter-class distances further. Since it is functionally based on the margins between the hardest positive distances and the hardest negative distances, the abnormal data will influence the training process enormously. Center Loss [21], first proposed for Face Recognition (FR), can learn centralized supervising signals for deep features of each identity, also helping improve the discrimination and integrity of representations. However, the generated signals with Center Loss [21] are unsuitable for the person Re-ID task. Firstly, compared with FR, collecting cross camera identity data is a tedious process, and consequently many identities in Re-ID datasets [9, 22, 7, 23, 24] are with deficient samples. Therefore, there is no guarantee for the representative power of the accumulated supervising signals. Secondly, compared with faces, the characteristics of pedestrians are much more unstable even within the same camera view across different time steps. Giving a universal signal for each identity is not acceptable.

To tackle the above problems, we propose a novel end-to-end Hierarchical Mining Network (HMN) for video-based person Re-ID from temporal learning and metric learning aspects. The proposed network is constructed by an Attentive Temporal Module (ATM) and a Dynamic Supervising Branch (DSB) with a Balancing Triplet Loss (BTL) for assisting the training. More concretely, the ATM consists of a Gated Recurrent Unit (GRU) [25] and a Squeeze and Excitation (SE) block [26]. Rather than generating frame-level attention weights, we fully explore the temporal information to assess the significance of each extracted feature (each channel). In this way, the discrete characteristics among frames can be better aggregated, and meanwhile the noise can be reduced enormously. Then, inspired by the teacher-student network [27, 28], we propose the DSB, which can generate supervising signals with high integrity and discrimination by constraining the intra-class consistency, to assist the training of the student branch. Specifically, we sample representations of the same identity into a graph and adopt the multi-head Graph Convolutional Network (GCN) [29] for feature updating in each mini-batch. Moreover, we also propose the Balancing Triplet Loss (BTL) to assist the training of the DSB, aiming at improving the persistence of metric learning and generating the output with larger inter-class variations and smaller intra-class variations. By establishing multiple supervision, the knowledge of the DSB can be distilled into the student branch, guaranteeing the student branch with better feature extraction capacity through single sequences. In the inference phase, the query and gallery sequences are without identity labels. Thus, the proposed DSB is unable to update representations by associating with the neighbors. Therefore, the DSB will be discarded.

Our contributions are summarized as follows:

- We propose a novel end-to-end Hierarchical Mining Network (HMN), which can generate representations of sequences with high integrity and discrimination.
- We propose a new Attentive Temporal Module (ATM), which is a combination of RNNs and the attention mechanism. ATM, with pedestrian perceiving capacity, adaptively evaluates the significance of each extracted feature (each channel) and assembles all the valuable ones together.
- We propose a new Dynamic Supervising Branch (DSB) to generate supervising signals by referring to intra-class samples within each mini-batch. It reduces the dependency on the number of intra-class samples and increases the discrimination of the signals greatly.
- We propose a new Balancing Triplet Loss (BTL) to increase the inter-class variations and reduce the intra-class variations of supervising signals from DSB. It further guarantees the dependability of the signals.

The superiority of our HMN over other state-of-the-arts is verified on two challenging video Re-ID datasets, MARS [9] and DukeMTMC-VideoReID [22]. Moreover, extensive experiments on image-based datasets, including Market1501 [7], DukeMTMC-ReID [23] and MSMT17 [24], further confirm the effectiveness of DSB and BTL.

II. Related Work

Existing person Re-ID works can be grouped into image-based Re-ID and video-based Re-ID. Compared with single images, video sequences possess additional temporal cues and also abundant appearance information of pedestrians, which can be utilized to generate more discriminative representations. In this section, we briefly review current temporal learning methods, deep metric learning in Re-ID tasks, and teacher-student networks.

A. Temporal Learning

1) Temporal pooling operations: Temporal pooling operations are widely adopted to aggregate features across different time steps. For example, Zheng *et al.* [9] apply *Max Pooling* and *Average Pooling* operations to aggregating frame-level features; Wang *et al.* [10] adopt *Average Pooling* operation on the extracted multi-scale features. Besides, unsupervised video Re-ID methods [30, 22] also utilize the *Average Pooling* operation. Attention mechanism, which has been proven very effective in classification [26, 31] and natural language

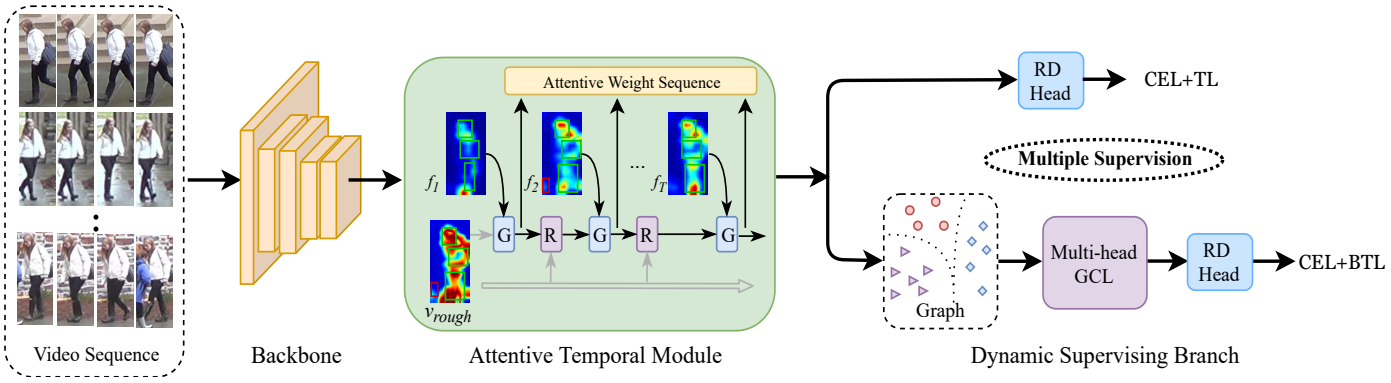


Fig. 2. Hierarchical Mining Network (HMN) for video-based person Re-ID. Given input video sequences, the Backbone Network extracts the spatial features of each frame firstly. The extracted features are fed into the Attentive Temporal Module (ATM) for further characteristics mining by referring to the temporal cues. Here, the block G and R refer to the Gated Recurrent Unit [25] and the recalibration block, respectively. (For better understanding, visually we transfer the activate distributions from channel dimension to spatial dimensions.) Then, the Dynamic Supervising Branch (DSB) dynamically generates supervising signals for each sample by referring to its neighbor samples. Loss functions are applied to the two branches individually. Here, CEL, TL, and BTL denote the Cross Entropy Loss, Batch-hard Triplet Loss [20], and the proposed Balancing Triplet Loss, respectively. Since in the inference no additional identity information of sequences is provided, the DSB is discarded, making HMN more concise and flexible.

processing [32], is also utilized for the temporal aggregation. For example, Li *et al.* [17] learn multiple spatial attention models to generate weights for part cues; Li *et al.* [19] exploit the multi scale temporal (long-term and short term) cues in the video sequences for generating proper attention weights for each frame-level feature.

2) Optical flow: Optical flow presents the pattern of apparent motion of objects. Many works introduce optical flow into video-based Re-ID to learn the temporal features. For instance, Simonyan *et al.* [33] learn the spatial feature and the temporal feature from the stacked optical flow by constructing a two-stream network; Chung *et al.* [34] present a two-stream architecture for appearance and temporal information learning, where an optical flow image along with the YUV image comprises the input to the deep learning network; Mclaughlin *et al.* [8] exploit long and short-term temporal cues by feeding the optical flow into the RNNs.

3) Recurrent Neural Network: Due to their abilities to learn long-term dependencies, RNNs are also adopted for video feature learning. Mclaughlin *et al.* [8] first extract frame-level features and use RNNs to model the temporal cues, and then aggregate frame-level features with simple pooling operations. Liu *et al.* [35] propose the refining recurrent unit to recover the missing characteristics and suppress the noise by referring to the temporal information. Cheng *et al.* [6] introduce the LSTM to learn attention weights for the feature snippets, enabling the resulting embeddings to be less affected by noisy frames. Bargal *et al.* [36] devise a formulation to simultaneously ground evidence in space and time for video action recognition and video captioning. Ramanishka *et al.* [37] propose Caption-Guided Visual Saliency to expose the region-to-word mapping in modern encoder-decoder networks and can provide accurate saliency heatmaps.

In a nutshell, current temporal learning methods still regard the features of each frame as an independent unit while aggregating. Since the valuable characteristics and the distracting factors are discretely distributed in frames, the integrity of the representations after aggregation will be inevitably weakened. In our proposed method, by regarding each feature as an independent individual and generating an attentive weight sequence for the feature sequence, the integrity of the final representations can be greatly improved.

B. Deep Metric Learning

In the Re-ID community, apart from the classification loss (*e.g.*, Cross-Entropy Loss, Label Smoothing Cross-Entropy Loss [38]), deep metric learning also works in the forms of various loss functions. Center Loss [21] learns a center for deep features of each class and penalizes the distance between a feature and its corresponding center. The triplet loss [20] optimizes the embedding space such that data points with the same identity are closer to each other than those with different identities. The function works conditioned on the distance gaps between the hardest negative pairs and the hardest positive pairs in the mini-batch and the predefined margin. Wang *et al.* [39] propose the Online Soft Mining (OSM), which assigns a continuous score rather than a binary one (dropping or keeping) to each sample to make full use of all the samples in the mini-batch during mining and introduces the attention mechanism to largely shrink the attention paid to the abnormal samples. Chen *et al.* [40] design a quadruplet loss to enforce a larger inter-class variation and a smaller intra-class variation compared with Triplet Loss [20].

C. Teacher-Student Network

Among teacher-student learning strategies, two classic ones are the knowledge distillation [27] and deep mutual learning [28]. Knowledge distillation aims to transfer the knowledge from a teacher network to a student network. During the learning process, a deep static pre-defined teacher is employed to guide the optimization of the student network. The shortcoming of this strategy is that a superior pre-trained model is required, which increases the complexity and workload of training. Comparatively, in deep mutual learning, rather than a one-way transfer between the teacher and student branches,

an ensemble of students learn collaboratively and teach each other throughout the training process. The two branches can be the same or different, but both of them will gain performance improvements after training. In unsupervised person Re-ID, Ge *et al.* [41] and Zhai *et al.* [42] also adopt teacher-student learning strategy to learn the knowledge cross domains.

Currently, only a few Re-ID methods adopt the teacher-student learning strategy. In our proposed method, we construct an individual branch that generates supervising signals for each category to conduct deep metric learning.

III. Methodology

To tackle the challenges brought by the inconsecutive characteristics of interested pedestrians in video sequences, we propose the Hierarchical Mining Network (HMN). An illustration of its architecture is shown in Fig. 2. Our HMN is constructed by a Backbone Network, an Attentive Temporal Module (ATM), and a Dynamic Supervising Branch (DSB). Besides, a Balancing Triplet Loss (BTL) is presented and applied for promoting the learning ability of the DSB. With these novel designs, the proposed network is able to learn discriminative representations with high characteristic integrity through mining the temporal cues of the sequences as well as the rich intra-class information. In the following subsections, we will introduce each component in detail.

A. Backbone Network

Following existing works like [6, 17, 19, 16], we adopt the pre-trained ResNet-50 [43] as the backbone for extracting the spatial features of each frame. Specifically, ResNet-50 [43] is constructed by one convolutional block termed as $conv1$ and four residual blocks termed as $conv2 \sim 5$, one global average pooling (GAP), and one 1000-dimensional fully connected (FC) layer. We make two modifications on the original network: 1) the last spatial down-sampling of the $conv5$ is removed for increasing the granularity of the features; 2) the final FC layer, which is for category classification, is discarded.

During training, we choose P identities each with K clips, making a total of PK clips in a mini-batch. For each video clip, there are T frames that are randomly sampled from the corresponding video tracklet. After the video clips are fed into the backbone network, the frame-level features and the feature sequence are obtained, denoted as $f_t \in \mathbb{R}^{2048}$ and $v = \{f_t\}_{t=1}^{T}$, respectively.

B. Attentive Temporal Module

Our Attentive Temporal Module (ATM) is a union of Recurrent Neural Networks (*i.e.*, RNN, Long Short-Term Memory (LSTM) [12], Gated Recurrent Unit (GRU) [25]) and the attention mechanism. See Fig. 3 for an illustration of its structure. It functionally simulates the mechanism of human brain in that the rough memory about objects is becoming increasingly concrete and integrated through multiple attentive observations. The RNNs, which are capable of learning long-term dependencies, are dedicated to perceiving the valuable

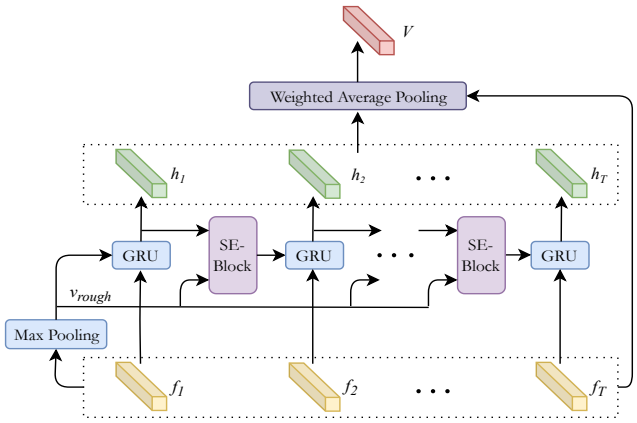


Fig. 3. The structure of the Attentive Temporal Module (ATM). It takes the feature sequence v as input and outputs the aggregated representation V. Since the module has the capacity of perceiving the pedestrians and temporal aggregation, the outputs are always with high integrity and discrimination.

characteristics through temporal analysis, and the attention mechanism aims at recalibrating the extracted features along the channel dimension. In this manner, the two mechanisms can complement each other with high compatibility, so as to mutually enhance each other for powerful representation learning. Intuitively, their combination, which has high cognition about characteristics of the interested persons in the video sequences, can relieve the burden of the backbone network and bring greater robustness to the learned model.

More specifically, the input to the ATM module is the feature sequence $v \in \mathbb{R}^{T\times 2048}$ from the backbone network. Firstly, we conduct the *Max Pooling* operation along the temporal dimension to get a rough representation v_{rough}. As each channel of the feature maps is considered as a feature detector [44], empirically, the rough *Max Pooling* operation is able to produce representations that cover both valuable and contaminated characteristics. Directly conducting the recalibration operation on v_{rough} is difficult to achieve the desired effect. Thus, we perform recalibration with the collaboration of temporal cues. Here, we feed the feature sequence into the GRU [25], which achieves similar performance to LSTM [12] with cheaper computation, to get the knowledge of the current time step. The GRU can be formulated as

$$\begin{aligned} r_t &= \sigma(W_r x_t + U_r h_{t-1}) \\ z_t &= \sigma(W_z x_t + U_z h_{t-1}) \\ \tilde{h}_t &= \tanh\left(W x_t + U(r_t \cdot h_{t-1})\right) \\ h_t &= (1 - z_t) h_{t-1} + z_t \tilde{h}_t, \end{aligned} \tag{1}$$

where r, $\tilde{h}$, z, and h stand for the reset gate, candidate activation, update gate, and activation, respectively. Rather than transmitting the extracted hidden states (activation) successively [6, 8], we utilize the v_{rough} and the activation together for generating temporal prior knowledge for the next time step. Therefore, we introduce the Squeeze and Excitation (SE) block [26] for the integration. Specifically, we squeeze and excite the last activation h_{t-1} to generate the recalibrating vector and take the multiplication of the vector and v_{rough} as

the new prior knowledge $h_{t-1}^{'}$. The process can be formulated as below:

$$h_{t-1}^{'} = v_{rough} \cdot (\sigma(W_2\delta(W_1 h_{t-1}))), \tag{2}$$

where $W_1 \in R^{\frac{C}{r}\times C}$, $W_2 \in R^{C\times\frac{C}{r}}$, and σ, δ refer to the $Sigmoid$ and $ReLU$, respectively. For the first time step with no prior knowledge provided, we straightforwardly take the v_{rough} as the latest activation. In this way, the prior knowledge of each step is not limited to the former frame, guaranteeing better generalization of the GRU. Besides, it can also maintain the consistency between the activation and the input feature sequence. After getting the original activation of all time steps, we put them together and form the activation sequence $H = \{h_t\}_{t=1}^{T}$. Then, we perform the *Softmax* operation on the collected activation sequence H along the temporal dimension to obtain the attentive weight sequence $A = \{a_t\}_{t=1}^{T}$ as follows:

$$a_{t,i} = \frac{e^{h_{t,i}}}{\sum_{t=0}^{T} e^{h_{t,i}}}, \tag{3}$$

where i represents the i^{th} channel of the activation, t is the frame number, and T is the sequence length. In the end, we conduct the *Weighted Average Pooling* operation through the sequence A on the sequence v to produce the aggregated features (denoted as V):

$$V = \frac{1}{T}\sum_{t=1}^{T} a_t \cdot f_t. \tag{4}$$

In this way, the proposed ATM can better preserve the valuable characteristics and eliminate the contaminated ones in the sequences, guaranteeing higher discrimination and integrity of the features.

C. Dynamic Supervising Branch

Center Loss [21] has been proven effective for Face Recognition (FR). It generates supervising signals for each identity by accumulating features of the corresponding samples. Although there is a momentum to control the learning rate of the centers, when the samples of some categories are deficient, the generated supervising signals tend to be unstable. Also, the video-based ReID task tacked in this work differs from FR in that the characteristics of pedestrians lack consistencies at different time steps due to the variations of occlusions, postures, and viewpoints. Thus, giving a universal supervising signal for each category (identity) is not suitable for the Re-ID task. Here, we propose a Dynamic Supervising Branch as a better alternative.

Specifically, our goal is to generate a specific discriminative supervising signal for each sample dynamically, relying on a small quantity of the intra-class samples. Rather than updating the signals with respect to the entire training set, we propose to simplify the process by basing it on each mini-batch. To this end, we employ the Graph Convolutional Network (GCN), which can refine and recalibrate each node's feature by referring to its neighbors. After obtaining the features from ATM over each mini-batch, we regroup them into P graphs (P corresponds to the number of identities), each with K nodes (K corresponds to the number of each identity's instances), and the i^{th} graph can be represented as $G^i = \{V_k^i\}_{k=1}^{K}$. Inspired by [6], we reform each graph into M small graphs by partitioning features into M parts along the channel dimension and the original graph G^i can be represented by $G^i = \{G_m^i\}_{m=1}^{M}$, where m is the small graph index. Since each small graph will be analyzed independently, the hard partition operation can straightforwardly reduce the impact of local noise on the global feature. Subsequently, the M small graphs (nodes along with the features) are fed into M individual GCNs for further aggregation and recalibration. The function of each small graph can be formulated as

$$G_m^{i'} = f(G_m^i, \hat{A}) = ReLU(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}G_m^i W_m), \tag{5}$$

where $\hat{A}$ is a matrix describing the graph structure through the cosine distance, $\hat{D}$ is the diagonal node degree matrix of $\hat{A}$, and W_m is a linear transfer matrix. After the refinement on all the graphs, we concatenate the split features of each sample along the channel dimension, denoted as V'.

Here, we take the upper branch and DSB as the student branch and the teacher branch, respectively. Features from the two branches will go through the parameter-shared Reduction Head (RD Head) and the classification layer. The shared layers guarantee the feature consistencies between the two branches. Specifically, the RD Head is constructed by a fully connected (FC) layer, Batch Normalization [45], and ReLU activation function. Here, we set the reduction rate of the RD Head as 2, compressing the feature dimension from 2048 to 1024.

The following supervision is applied to both features and probability aspects. In the feature aspect, we take the output of the teacher branch as the supervising signal and calculate the mean square error (MSE) loss between the features of the two branches. The function can be formulated as

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{C}(V_i - V_i^{'})^2, \tag{6}$$

where $V^{'}$ and V represent the features from the teacher branch and the student branch, respectively, i is the channel index, and C is the length of the channel dimension. From the probability aspect, the Kullback-Leibler (KL) divergence, which measures the probability distribution differences, is adopted for fitting the two branches. The output of the teacher branch is also taken as the learning target of the student branch, the same as the previous operation. We firstly conduct the *Softmax* operation on the final classification vectors for normalization. Then KL divergence Loss is calculated for the normalized vectors. The function can be formulated as

$$D_{KL}(P_s||P_t) = -\sum_{x\in X} P_s(x)\log(\frac{P_t(x)}{P_s(x)}), \tag{7}$$

where P_s and P_t represent the ID probability of student and teacher branches, respectively, and X is the probability space.

Compared with Center Loss [21], the proposed DSB greatly reduces the dependencies upon the number of intra-class samples when generating supervising signals. Since the generated signals are further supervised by the loss functions, they are

more reliable and stable than Center Loss generated ones. In the inference phase, we have no identity labels of sequences, which means that the sampling and feature updating are unacceptable. Therefore, DSB will be discarded, making our network more concise and feasible at the inference stage.

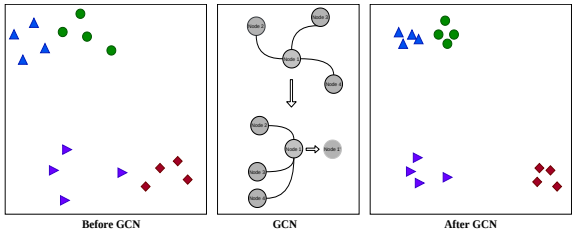


Fig. 4. Feature distributions in each mini-batch before and after the multi-head GCNs. Here, different symbols represent different identities. After the GCNs' refinement, the intra-class distances are smaller than inter-class distances. Since traditional Triplet Loss works conditioned on the distance margin between the hardest negative and the hardest positive samples, it lacks persistence and loses the constraint on metric learning.

D. Loss Function

1) Balancing Triplet Loss: Batch-hard Triplet Loss (TL) [20] has been widely used in both image-based and video-based person Re-ID methods due to its great power of metric learning. The function works conditioned on the margin between the hardest positive distances and the hardest negative distances. In the DSB, due to the refinement of GCNs on the features, the magnitude of intra-class distances is much smaller than that of the inter-class distances (as shown in Fig. 4). The functional basis of the TL then collapses and as a result, the loss function lacks persistence on the DSB. Therefore, we propose a Balancing Triplet Loss (BTL) to improve the adaptability of TL by adjusting the distance magnitude of the hardest pairs.

To address the limitation of TL, a straightforward way is to multiply a hard attenuated parameter with the hardest negative distances. However, different parameters have different influences on the loss function and the distance magnitude of the hardest pairs varies with the training processing, a hard parameter is not acceptable. Therefore, we propose to generate a soft parameter. In particular, in each mini-batch, we take the average division value of the hardest positive distances and the hardest negative distances of samples in the mini-batch as the soft attenuated parameter, denoted as β. The function can be formulated as

$$\beta = \frac{1}{PK}\sum_{i=0}^{PK}\frac{D(x_a^i, x_p^i)}{D(x_a^i, x_n^i)}, \tag{8}$$

where x_a, x_p, x_n, i, and $D()$ are the anchor, the hardest positive sample, the hardest negative sample, sequence index, and the distance measure function, respectively. Besides, we pre-define the lower boundary γ to avoid the embedding space distortion caused by over enlarging the inter-class distances. By multiplying the β with the hardest negative distances, the two distances can be pulled to the same magnitude level dynamically. The final function can be formulated as

$$\begin{aligned} L_{\text{BTL}} = \sum_{i=1}^{P}\sum_{a=1}^{K}\Big[m + \max_{p=1,\dots,K} D(x_a^{(i)}, x_p^{(i)}) \\ - \min_{\substack{j=1,\dots,P \\ n=1,\dots,K \\ j\neq i}} \beta D(x_a^{(i)}, x_n^{(j)})\Big]_+. \end{aligned} \tag{9}$$

In this paper, we set the boundary γ as 0.1.

2) Overall Loss: As shown in Fig. 2, each branch has its own independent loss functions for training. For the student branch, we calculate the Cross Entropy Loss (CEL) and Batch-hard Triplet Loss (TL) to predict the pedestrians' ID and learn the representations, respectively. For the teacher branch (DSB), we replace the TL with BTL to increase the persistence of the teacher branch's metric learning. As mentioned, we adopt MSE loss and KL divergence to enhance the supervisions on the student branch. The overall loss can then be formulated as

$$\begin{aligned} L_{final} = L_{CEL_S} + L_{TL_S} + L_{CEL_T} \\ + L_{BTL_T} + \lambda(D_{KL}(P||Q) + L_{MSE}), \end{aligned} \tag{10}$$

where λ is a predefined hyperparameter for balancing the supervisions, and the capitals S, T refer to losses in the teacher branch and the student branch, respectively. Here, we set the λ as 0.4. Following the instructions of [20], we set the margin of TL as 0.3. As for BTL, considering the attenuated parameter β would weaken the distance gaps, we set the margin as 0.1. The network is trained end-to-end by jointly optimizing all the loss functions.

IV. Experiments

A. Datasets

In this paper, we adopt the video-based Re-ID datasets for evaluating the performance of the proposed network. Besides, image-based Re-ID datasets are also adopted to further verify the effectiveness of the Dynamic Supervising Branch (DSB) and Balancing Triplet Loss (BTL) specifically.

1) Video-based Datasets: MARS [9] dataset is one of the largest published video-based person Re-ID datasets. It contains 1,261 pedestrians with each captured by at least two cameras, out of six cameras in total. There are 20,715 sequences, containing 3,248 distracting tracklets due to false detection or tracking. Every identity in the training set has 13 video tracklets on average, and each tracklet has 59 frames on average. *DukeMTMC-VideoReID* [22] is another large-scale video-based person Re-ID dataset. It is captured in outdoor scenes with noisy background and suffers from occlusions and variations in illumination, poses and viewpoints. The dataset is composed of 4,832 tracklets from 1,812 identities. It is split into 702, 702, and 408 identities for training, testing, and distracting, respectively. The training and testing sets contain 369,656 frames of 2,196 tracklets and 445,764 frames of 2,636 tracklets, respectively. Each tracklet has 168 frames on average. Since the bounding boxes are all annotated manually rather than using algorithms, false detection samples are much less than MARS [9] dataset.

We follow the protocol of MARS [9] and DukeMTMC-VideoReID [22] to split the training and testing sets, guaranteeing that no overlapping identities exist.

2) Image-based Datasets: Market1501 [7] collects a total of 12,936 training images of 751 identities across six cameras without overlapping views As for the testing data, gallery and query sets consist of 19,732 and 3,368 images respectively with 750 different identities. *DukeMTMC-ReID* [23] includes 36,411 labeled images from 1,404 identities. 702 identities are randomly selected for training and the rest are used for testing. There are 16,522 training images, 2,228 query images, and 17661 gallery images, respectively. *MSMT17* [24] is the most challenging and large-scale dataset consisting of 126,441 bounding boxes of 4,101 identities captured by 15 cameras. In this dataset, 32,621 images of 1,041 identities are split for training, and the rest are used for testing.

B. Evaluation Metrics

Since person Re-ID lies in between image classification and instance retrieval [46] in terms of the relationship between the training and the testing set (none overlapping identities exist), we employ both the Cumulative Marching Characteristics (CMC) curve and mean Average Precision (mAP) for evaluation. CMC curve is used for evaluating the accuracy of the person retrieval. For each query, its Average Precision (AP) is calculated from its precision-recall curve. The mAP is the mean value of AP across all queries.

C. Implementation Details

Our experiments are implemented on Pytorch platform and with one Nvidia TiTan RTX GPU (24GB memory size). All the input images are resized to 256×128. In the training phase, we randomly select 8 frames from every video clip and group them into a tracklet. Each frame is augmented by random horizontal flipping, normalization, and random crop. As for random crop (with the probability of 50%), we first enlarge width and height by 1/8 and then crop it to the original size. Each mini-batch contains exactly 8 identities in total, each with 4 individual sequences. For the identities with less than 4 sequences, we randomly duplicate the existing sequences of the corresponding identity, to ensure every ID has 4 sequences. Adam with weight decay 0.0005 is adopted for optimization. The learning rate is initialized as 0.0003 and decreased by $\times$ 0.1 per 200 epochs. We train the models for 800 epochs. In the inference phase, we chop each video (containing N frames) into $N/8$ tracklets firstly, and then average the extracted features of all the tracklets as the final representation.

D. Ablation Study

1) Effects of Components: To verify the effectiveness of each component in the Hierarchical Mining Network (HMN), we conduct several analytic experiments on MARS [9] and DukeMTMC-VideoReID [22] datasets. The results are summarized in TABLE I. The 'Baseline(B)' is the combination of the modified ResNet-50 backbone [43] and the Reduction Head, optimized by the Cross Entropy Loss. 'B+SE', 'B+GRU', and 'B+ATM(SE)' refer to adding the Squeeze and Excitation block [26], the Gated Recurrent Unit [25], and the proposed Attentive Temporal Module on the 'Baseline(B)', respectively. By comparing the 'B+ATM(SE)' with 'B+GRU' and 'B+SE', we observe the Rank-1 and mAP are improved by at least 1.1% and 0.7% on MARS dataset and 0.7% and 0.4% on DukeMTMC-VideoReID dataset. These results prove that our proposed ATM perfectly inherits the advantages of both the GRU and the attention mechanism and improves the discrimination of features through temporal mining. Besides, to further validate the proposed ATM, we modify the original one by replacing the SE block by a simple residual block which is constructed by two linear layers and one ReLU activation. We name it ATM(Res). By comparing the performance, the two kinds of ATMs achieve similar accuracies, demonstrating the rationality of the overall design. Here, 'B+Center' is a combination of 'Baseline' and Center Loss [21]. 'B+DSB(TL)' and 'B+DSB(BTL)' are the DSB optimized by the Triplet Loss and the proposed Balancing Triplet Loss, respectively. By comparing 'B+DSB(BTL)' with 'B+Center' and 'B+DSB(TL)', we can see the Rank-1 and mAP are also get enhanced on the two datasets, especially the mAP. We also test hard attenuated parameter β, such as 0.1 and 0.05, in TL. The hard βs lead to worse performance than TL because of destroying the balance between the distances. Therefore, it is validated that our proposed DSB combining with BTL can generate more representative supervising signals for samples. At last, the 'HMN' can achieve the best performance compared with others, demonstrating that each component can work individually and cooperatively.

Besides, to further evaluate the DSB and BTL, we also conduct experiments on current image-based Re-ID datasets, *i.e.*, Market1501 [7], DukeMTMC-ReID [23], MSMT17 [24]. The results are shown in TABLE II. As can be seen, our DSB achieves similar accuracy compared with Center Loss [21]. The main reason is that compared with video sequences, the single images contain fewer variations of pedestrians. Therefore, only adopting single images in each mini-batch is hard to generate representative enough supervising signals. We also compare our network with the state-of-the-art method SCSN [47]. There are still gaps between the algorithm that adopts advanced network architecture.

2) Effect of Multi-head GCN: In the DSB, we divide the global features into N parts along the channel dimension and feed them into the Multi-head GCN to learn supervising signals with good discrimination. Here we conduct experiments to verify the effectiveness of the multi-head GCN. In these experiments, we set N as 1, 2, 4, and 8, respectively, to find the most suitable one for the video Re-ID task. As shown in TABLE III, when $N = 2$ and $N = 4$, the model can achieve the highest Rank-1 accuracy and mAP accuracy, respectively. When N is larger than 4, both of the Rank-1 and mAP accuracy drop. This can be explained as follows: with the increasing number of partitions, the impact of noise on the corresponding portion will also increase, indirectly aggravating the contamination of the final global representation.

3) Computation and Parameters of the Model: In TABLE IV, we present the computation and parameters of the models.

TABLE I
ABLATION STUDY ON MARS AND DUKEMTMC-VIDEOREID DATASETS. CMC CURVE AND MAP ARE PRESENTED FOR DEMONSTRATING THE EFFECTIVENESS OF EACH COMPONENT.

	MARS				DukeMTMC-VideoReID			
Model	Rank-1	Rank-5	Rank-20	mAP	Rank-1	Rank-5	Rank-20	mAP
Baseline(B)	84.6	94.9	98.0	79.1	93.6	99.0	99.7	93.2
B+GRU	85.7	95.1	97.6	79.0	94.7	99.1	99.3	93.3
B+SE	86.0	95.2	97.7	80.5	94.6	99.1	99.7	93.2
B+ATM(RES)	86.6	95.0	97.9	80.8	94.8	99.3	99.7	93.6
B+ATM(SE)	87.1	95.8	97.9	81.2	95.3	99.2	99.7	93.6
B+Center	86.8	95.4	97.7	80.4	94.9	99.0	99.6	93.5
B+DSB(TL β=0.1)	86.7	94.7	97.6	80.7	95.0	99.0	99.6	93.6
B+DSB(TL β=0.05)	86.6	95.7	98.1	81.0	95.3	98.9	99.6	93.8
B+DSB(TL)	87.0	95.6	97.9	81.1	95.1	98.7	99.7	94.3
B+DSB(BTL)	87.5	95.8	**98.1**	81.5	95.5	**99.6**	99.7	94.8
HMN	**88.5**	**96.2**	**98.1**	**82.6**	**96.3**	99.2	**99.8**	**95.1**

TABLE II
EXPERIMENTS CONDUCTED ON IMAGE-BASED RE-ID DATASETS, INCLUDING MARKET1501, DUKEMTMC-REID, AND MSMT17, TO VERIFY THE EFFECTIVENESS OF THE DSB AND BTL.

	Market1501				DukeMTMC-ReID				MSMT17			
Model	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
BaseLine	91.5	97.1	98.2	80.6	83.3	91.8	94.7	70.8	69.2	82.5	86.8	45.0
+Center	92.8	**97.7**	**98.6**	**84.0**	85.9	93.1	95.2	74.1	73.6	85.1	88.7	**48.6**
+DSB(BTL)	**93.1**	97.4	**98.6**	83.3	**86.1**	**93.3**	**95.4**	**74.6**	**73.9**	**85.4**	**88.8**	48.3
SCSN [47]	**95.7**	-	-	**88.5**	**91.0**	-	-	**79.0**	**83.8**	-	-	**58.5**

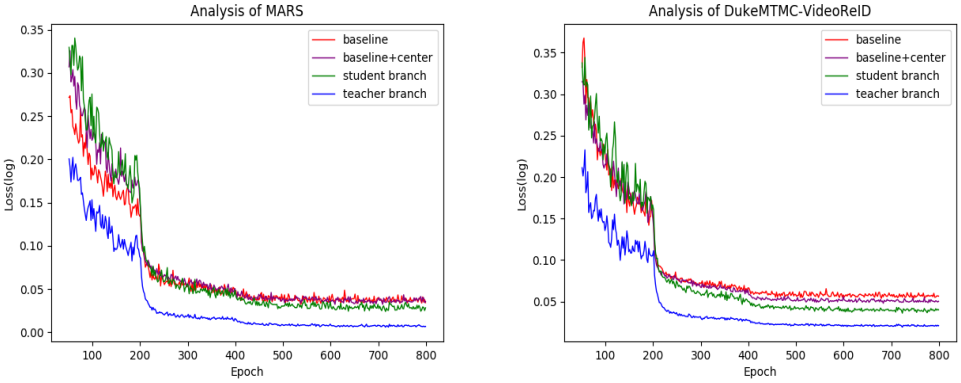


Fig. 5. Convergence comparison on MARS and DukeMTMC-VideoReID datasets. The curves represent the final Cross Entropy Losses of different models. For better visualization, we present losses from the 50th epoch and *log* function is applied.

TABLE III
ANALYSIS OF THE NUMBER OF GCN HEADS ON MARS DATASET. THE EXPERIMENTS ARE BASED ON THE HMN.

	MARS			
Model	Rank-1	Rank-5	Rank-20	mAP
N=1	87.9	96.0	**98.3**	82.1
N=2	**88.5**	**96.2**	98.1	82.6
N=4	88.3	**96.2**	98.2	**83.0**
N=8	87.8	**96.2**	97.9	82.1

TABLE IV
COMPARISON BETWEEN LSTM, GRU, AND THE PROPOSED ATM ON PARAMETERS AND COMPUTATION COST.

	Parameters(M)	Flops(G)
LSTM	33.5708	0.2687
GRU	25.1781	0.2015
ATM	26.2267	0.2099

Specifically, we compare the LSTM [12], GRU [25], and the proposed ATM. As we can see, the LSTM is heavier than GRU and the ATM by a large margin, reaching 33.5708M parameters and 0.2687G flops. Adding a SE block [26] in the GRU only brings 1.0486M parameters and 0.0084G flops. In conclusion, the proposed ATM contains much less parameters and computates efficiently than the LSTM.

4) Convergence Analysis: To further verify the superiority of our proposed DSB, we show the convergence degree of various models on MARS and Dukemtmc-VideoReID datasets in Fig. 5. The Cross Entropy Loss, which serves as a metric to measure the distance between the model's predictions and labels, is adopted for evaluating the convergence of the 'Baseline', the 'Baseline+Center', and the 'Baseline+DSB'. For the 'Baseline+DSB', we visualize the losses of the teacher branch (DSB) and the student branch separately. As can be

seen, 200 epochs is the turning point for the three networks, before which models are quite unstable especially the student branch in 'Baseline+DSB'. We owe this to the same unstable supervising signals. After the turning point, the superiority of the teacher branch becomes apparent. It converges much faster and more stable than others, demonstrating its advantages in generating supervising signals. By comparing the 'Baseline+center' and the student branch, although their convergence speeds are parallel, the smaller loss value of the student branch demonstrates closer distances between the model predictions and the labels. To sum up, the proposed DSB can dynamically generate superior supervising signals over the accumulated ones of Center Loss [21].

TABLE V
COMPARISON OF THE HMN WITH OTHER STATE-OF-THE-ART METHODS ON MARS DATASETS, WITH RANK-1, -5, -20 ACCURACY AND MAP REPORTED. * AND + REFER TO THE RE-RANKING OPERATION AND THE ADDITION OF OPTICAL FLOW, RESPECTIVELY.

Dataset	MARS			
Rank@R	Rank-1	Rank-5	Rank-20	mAP
TAM+SRM [48]	70.6	90.0	97.6	50.7
SPL+DRL [49]	74.8	86.7	93.4	-
ETAP-Net[22]	80.75	92.07	96.11	67.39
MSTA [50]	82.28	94.32	97.60	69.42
DRSAN [17]	82.3	-	-	65.8
COSAM [18]	84.9	95.5	97.9	79.9
CSACSE^{+} [6]	86.3	94.7	98.2	76.1
A3D [51]	86.3	95.5	-	80.4
STA [16]	86.3	95.7	98.1	80.8
AMEM [52]	86.7	94.0	97.1	79.3
AFDTA [5]	87.0	95.4	98.7	78.2
GLTR [19]	87.02	95.76	98.23	78.47
SCAN [53]	87.2	95.2	98.1	77.2
FGRA [54]	87.3	96.0	98.1	81.2
VRSTC [55]	88.5	96.5	-	82.3
M3D [56]	88.63	96.41	98.77	79.46
VDK [57]	**89.4**	**96.8**	-	83.1
RQEN * [58]	77.83	88.84	94.29	71.14
STA* [16]	87.2	96.2	98.6	87.7
MSTA* [50]	84.08	93.52	98.00	79.67
M3D* [56]	88.87	96.64	**98.64**	85.46
HMN (Ours)	88.47 $\pm$ 0.20	96.00 $\pm$ 0.16	97.93 $\pm$ 0.21	82.37 $\pm$ 0.12
HMN* (Ours)	89.0 $\pm$ 0.08	96.17 $\pm$ 0.12	98.40 $\pm$ 0	**88.80** $\pm$ **0.16**

TABLE VI
COMPARISON OF THE HMN WITH OTHER STATE-OF-THE-ART METHODS ON DUKEMTMC-VIDEOREID DATASETS, WITH RANK-1, -5, -20 ACCURACY AND MAP REPORTED.

Dataset	DukeMTMC-VideoReID			
Rank@R	Rank-1	Rank-5	Rank-20	mAP
ETAP-Net[22]	83.62	94.59	97.58	78.34
VRSTC [55]	95.0	99.1	-	93.5
VKD [57]	95.2	98.6	-	93.5
COSAM [18]	95.4	99.3	**99.8**	94.1
M3D [56]	95.49	99.30	99.72	93.67
STA [16]	96.2	**99.3**	99.6	94.9
GLTR [19]	96.29	**99.30**	99.71	93.74
HMN (Ours)	**96.23** $\pm$ **0.04**	99.20 $\pm$ 0.08	**99.8** $\pm$ **0.08**	**95.13** $\pm$ **0.05**

E. Visualization of the Sequences' Activate Regions

Apart from the experiments in the accuracy aspect, we also visualize sequences' activate regions of the baseline and the proposed HMN in Fig. 6. The visualization is achieved by summing up the absolute value of the feature maps along the channel dimension. For the baseline, we directly visualize the feature maps of the backbone network. As for the HMN, the product of feature maps from the backbone network and the temporal attentive vectors from the ATM are visualized to present the advantages of the temporal analysis.

By comparing activate heat maps of the first sequence, we can find that the HMN covers more characteristics of the pedestrian, demonstrating the better feature extraction ability in scenes without distractors. By comparing the activate heat maps of the following four sequences in the challenging scenes, we can conclude that the HMN can better distinguish the various distractors, *e.g.*, umbrellas, cars, other pedestrians, that are framed out by the red boxes, and cover more characteristics of pedestrians. It demonstrates that the HMN can extract representations with higher integrity and discrimination in challenging scenes.

F. Similarity Comparison

In Fig. 7, we select several challenging sequences and compare the similarities through cosine distance. The two sides of each sequence are the similarities between the corresponding query sequence of the two networks. By comparing the similarities, we can find that 1: Representations from HMN have higher intra-class similarities. Even the distractors break the dynamics of certain parts of the body and lead to inconsecutive characteristics, the network can cover them well. 2: Representations from HMN have lower inter-class similarities. When other pedestrians are visually similar, the HMN can extract representations that have low similarities between queries. It demonstrates that the HMN also focuses on some details of pedestrians rather than just salient characteristics. In all, we can conclude that our proposed HMN can extract discriminative representations with high integrity that can perform well under both normal and challenging scenes.

G. Comparison with State-of-the-art Methods

We compare the performance of our proposed video-based Re-ID system with other state-of-the-art methods from the literature on MARS and DukeMTMC-VideoReID datasets in TABLE V and TABLE VI, respectively.

As shown in TABLE V, our method achieves 88.5% and 89.2% on the Rank-1 accuracy before and after re-ranking [59]. Two-stream M3D [56] outperforms our method by 0.13% on Rank-1 accuracy before re-ranking [59]. Their method uses a 3D convolution layer called the Multi-scale 3D convolution layer to learn spatial-temporal cues. Besides, the two-stream feature fusion architecture also plays an important role in its feature extraction. However, their mAP, before and after re-ranking, are 3.14% and 3.04% lower than ours, respectively. We attribute this to the contribution of our hierarchical mining strategy that helps improve our recall rate. VDK [57] also achieves the state-of-the-art results. This method also adopts the teacher-student framework to treat the visual variety as a supervision signal. However, compared with our teacher-student strategy, they adopt two independent networks for learning, which is much tedious than our DSB. Overall, we can

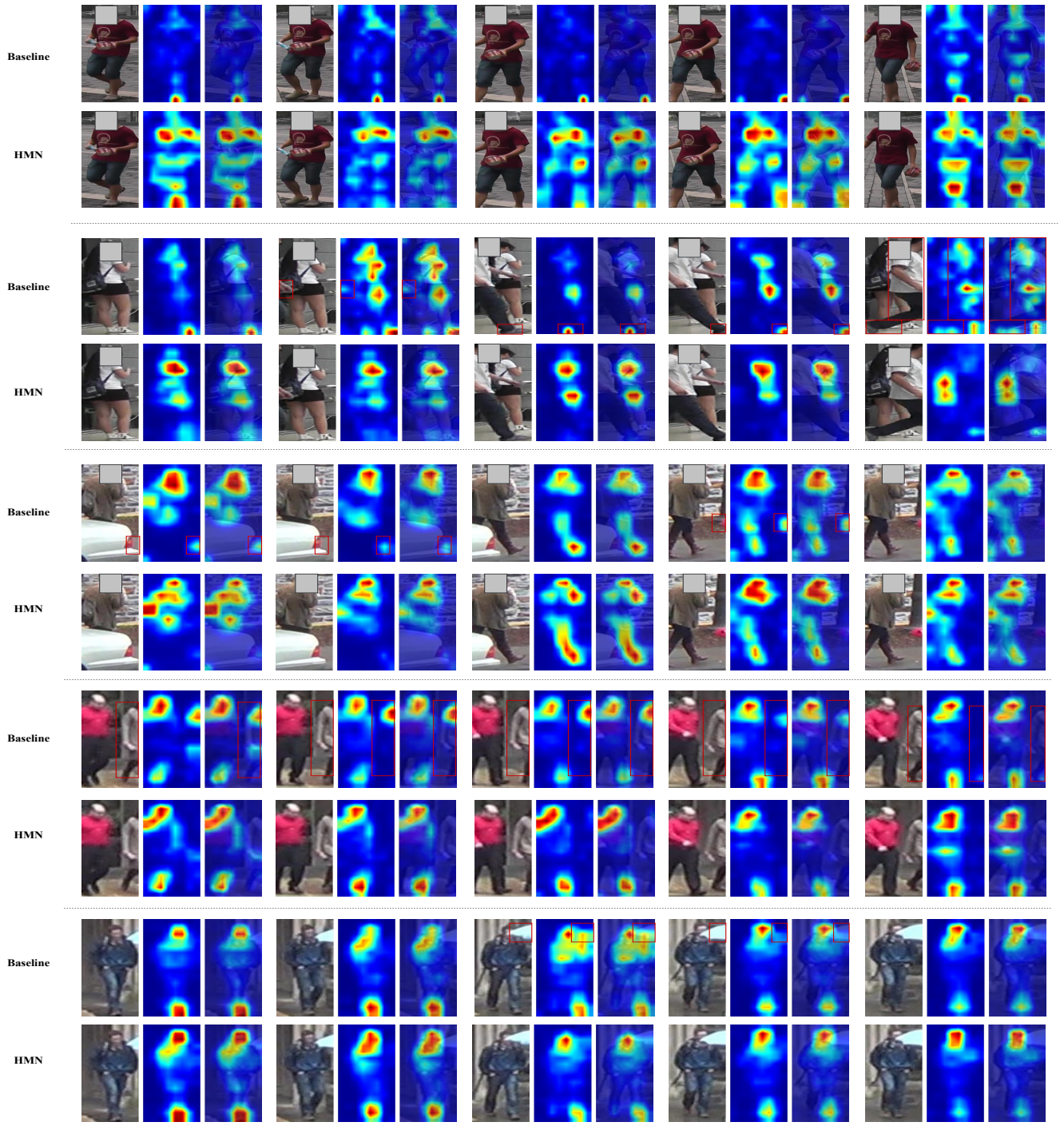


Fig. 6. Activate heat map visualization of the 'Baseline' and the proposed HMN. For the 'Baseline', we directly visualize the feature maps from the backbone network. As for the HMN, we visualize the multiplications between the temporal attentive vectors and the feature maps from the backbone network. By comparing the activate heat maps of the two networks, the proposed method can extract more pedestrians' characteristics. Besides, it is more robust to the various distracting factors.

achieve competitive CMC Rank-1, Rank-5, Rank-20 accuracy, and best mAP on the MARS dataset.

As shown in TABLE VI, our method achieves 96.3% and 95.1% on Rank-1 accuracy and mAP, respectively. STA [16] and GLTR [19] achieve higher Rank-5 accuracy than ours. Specifically, STA [16] fully exploits the discriminative parts of one target person in both spatial and temporal dimensions for generating a 2D attention score matrix, which measures the importance of spatial parts across different frames and conducts attentive aggregation. GLTR [19] firstly models the short-temporal temporal cues among adjacent frames, then captures the long-term relations among nonconsecutive frames, and finally generates the final representation by aggregating the two kinds of cues. However, on Rank-1, Rank-20, and mAP, our HMN can achieve much better performance than them.

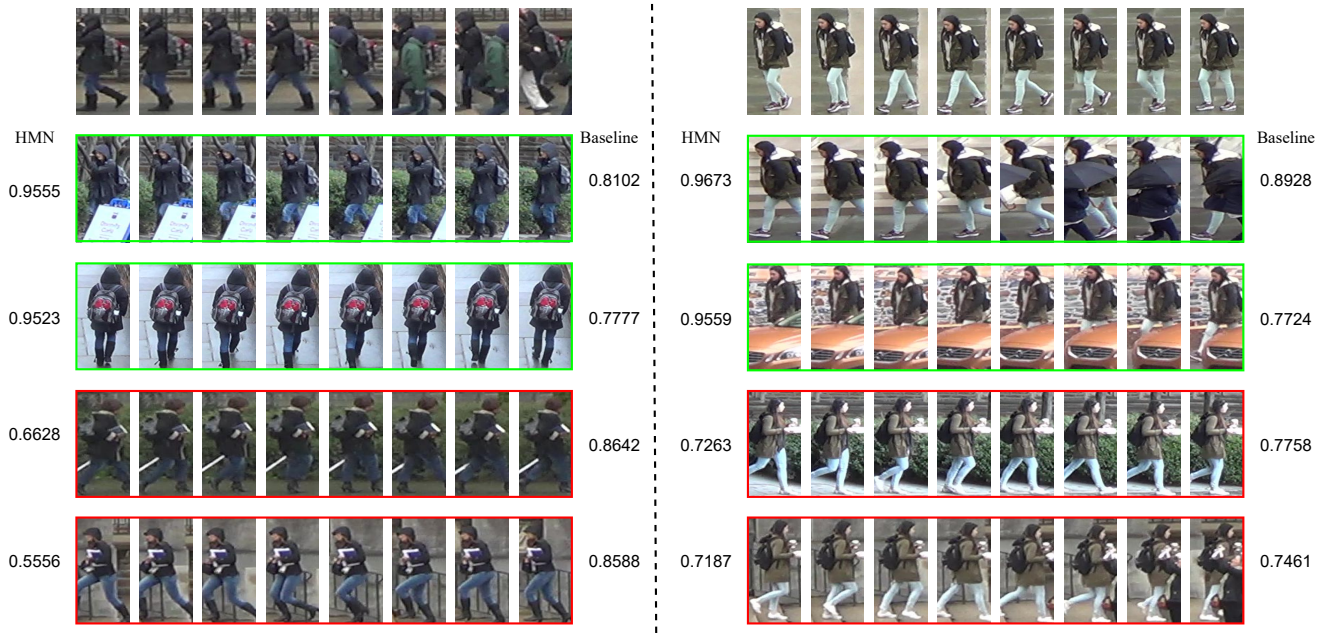


Fig. 7. Similarity comparison between query sequences and gallery sequences. The sequences in the first row are the queries. For each column, the right side and the left side of each sequence present the representation similarities of the HMN and baseline network, respectively.

By comparing results on the two large-scale video Re-ID datasets, our method can always achieve the highest mAP accuracy. These results demonstrate that our strategy, which mines discriminative features with high integrity by referring to the temporal and intra-class knowledge, is effective. Especially on the DukeMTMC-VideoReID dataset, which includes a huge amount of occlusions, both the mAP and CMC accuracy are much higher than those of other methods.

V. Conclusion

This paper proposes an innovative Hierarchical Mining Network (HMN) for video-based person Re-ID. The proposed HMN is capable of extracting discriminative representations with high integrity even over sequences where the characteristics of pedestrians are not consecutive. By incorporating the RNNs and attention mechanism, the proposed Attentive Temporal Module (ATM) evaluates each feature individual from the whole frame by referring to the temporal information. In this way, scattered characteristics of pedestrians in the sequences can be better aggregated. Furthermore, we propose a novel Dynamic Supervising Branch (DSB) along with the Balancing Triplet Loss (BTL) to generate supervising signals by fully exploring the intra-class samples in each mini-batch, reducing the dependency on the quantity of intra-class samples enormously. The experimental results on the standard benchmarks demonstrate the effectiveness of the proposed method, and the extensive ablation studies further verify the effect of each component in our network.

Acknowledgment

This research was supported partially by the National Key Research and Development Program of China (Grant Nos. 2018AAA0102702, 2016QY01W0200), the National Natural Science Foundation of China (Grant Nos. 61876146, 62006244, 62072354, 62036007, 62050175), the Fundamental Research Funds for the Central Universities (Grant No. JB210209), the Open Fund Project of Key Laboratory of Flight Technology and Flight Safety of Civil Aviation (FZ2020KF10).

References

[1] Xiaogang Wang. Intelligent multi-camera video surveillance: A review. *Pattern recognition letters*, 34(1):3–19, 2013.

[2] Chen Change Loy, Tao Xiang, and Shaogang Gong. Multi-camera activity correlation analysis. In *CVPR*, pages 1988–1995. IEEE, 2009.

[3] Shoou-I Yu, Yi Yang, and Alexander Hauptmann. Harry potter's marauder's map: Localizing and tracking multiple persons-of-interest by nonnegative discretization. In *CVPR*, pages 3714–3720, 2013.

[4] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele. Multiple people tracking by lifted multicut and person re-identification. In *CVPR*, pages 3539–3548, 2017.

[5] Yiru Zhao, Xu Shen, Zhongming Jin, Hongtao Lu, and Xian sheng Hua. Attribute-driven feature disentangling and temporal aggregation for video person re-identification. In *CVPR*, pages 4913–4922, 2019.

[6] Dapeng Chen, Hongsheng Li, Tong Xiao, Shuai Yi, and Xiaogang Wang. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In *CVPR*, pages 1169–1178, 2018.

[7] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, pages 1116–1124, 2015.

[8] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016.

[9] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, pages 868–884. Springer, 2016.

[10] Zhikang Wang, Lihuo He, Xinbo Gao, and Yuanfei Huang. Multi-scale spatial-temporal network for person re-identification. In *ICASSP*, pages 2052–2056. IEEE, 2019.

[11] Michael C Mozer. A focused back-propagation algorithm for temporal pattern recognition. *Complex systems*, 3(4):349–381, 1989.

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

[13] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, pages 4733–4742, 2017.

[14] Jinjie You, Ancong Wu, Xiang Li, and Wei-Shi Zheng. Top-push video-based person re-identification. In *CVPR*, pages 1345–1353, 2016.

[15] Hao Liu, Zequn Jie, Karlekar Jayashree, Meibin Qi, Jianguo Jiang, Shuicheng Yan, and Jiashi Feng. Video-based person re-identification with accumulative motion context. *tcsvt*, 28(10):2788–2802, 2017.

[16] Yang Fu, Xiaoyang Wang, Yunchao Wei, and Thomas Huang. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In *AAAI*. 2019.

[17] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for video-based person re-identification. In *CVPR*, pages 369–378, 2018.

[18] Arulkumar Subramaniam, Athira Nambiar, and Anurag Mittal. Co-segmentation inspired attention networks for video-based person re-identification. In *ICCV*, pages 562–572, 2019.

[19] Jianing Li, Jingdong Wang, Qi Tian, Wen Gao, and Shiliang Zhang. Global-local temporal representations for video person re-identification. In *ICCV*, pages 3958–3967, 2019.

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

[21] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *ECCV*, pages 499–515. Springer, 2016.

[22] Yu Wu, Yutian Lin, Xuanyi Dong, Yan Yan, Wanli Ouyang, and Yi Yang. Exploit the unknown gradually: One-shot video-based person re-identification by step-wise learning. In *CVPR*, pages 5177–5186, 2018.

[23] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, pages 3754–3762, 2017.

[24] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, pages 79–88, 2018.

[25] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[26] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018.

[27] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[28] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018.

[29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[30] Minxian Li, Xiatian Zhu, and Shaogang Gong. Unsupervised person re-identification by deep learning tracklet association. In *ECCV*, pages 737–753, 2018.

[31] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[33] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014.

[34] Dahjung Chung, Khalid Tahboub, and Edward J Delp. A two stream siamese convolutional neural network for person re-identification. In *ICCV*, pages 1983–1991, 2017.

[35] Yiheng Liu, Zhenxun Yuan, Wengang Zhou, and Houqiang Li. Spatial and temporal mutual promotion for video-based person re-identification. In *AAAI*, volume 33, pages 8786–8793, 2019.

[36] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. Excitation backprop for rnns. In *CVPR*, pages 1440–1449, 2018.

[37] Vasili Ramanishka, Abir Das, Jianming Zhang, and Kate Saenko. Top-down visual saliency guided by captions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7206–7215, 2017.

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.

[39] Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, and Neil M Robertson. Deep metric learning by online soft mining and class-aware attention. In *AAAI*,

volume 33, pages 5361–5368, 2019.

[40] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017.

[41] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. 2019.

[42] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. *ECCV*, 2020.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[44] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833. Springer, 2014.

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

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

[47] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. Salience-guided cascaded suppression network for person re-identification. In *CVPR*, pages 3300–3310, 2020.

[48] Zhen Zhou, Yan Huang, Wei Wang, Liang Wang, and Tieniu Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, pages 4747–4756, 2017.

[49] Deqiang Ouyang, Jie Shao, Yonghui Zhan, Yang Yang, and Heng Tao Shen. Video-based person re-identification via self-paced learning and deep reinforcement learning framework. In *ACM Multimedia*, pages 1562–1570. ACM, 2018.

[50] Wei Zhang, Xuanyu He, Xiaodong Yu, Weizhi Lu, Zhengjun Zha, and Qi Tian. A multi-scale spatial-temporal attention model for person re-identification in videos. *Transactions on Image Processing*, 29:3365–3373, 2019.

[51] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Learning recurrent 3d attention for video-based person re-identification. *Transactions on Image Processing*, 2020.

[52] Shuzhao Li, Huimin Yu, and Haoji Hu. Appearance and motion enhancement for video-based person re-identification. In *AAAI*, pages 11394–11401, 2020.

[53] Ruimao Zhang, Jingyu Li, Hongbin Sun, Yuying Ge, Ping Luo, Xiaogang Wang, and Liang Lin. Scan: Self-and-collaborative attention network for video person re-identification. *Transactions on Image Processing*, 28(10):4870–4882, 2019.

[54] Zengqun Chen, Zhiheng Zhou, Junchu Huang, Pengyu Zhang, and Bo Li. Frame-guided region-aligned representation for video person re-identification. In *AAAI*, pages 10591–10598, 2020.

[55] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Vrstc: Occlusion-free video person re-identification. In *CVPR*, pages 7183–7192, 2019.

[56] Jianing Li, Shiliang Zhang, and Tiejun Huang. Multi-scale temporal cues learning for video person re-identification. *Transactions on Image Processing*, 29:4461–4473, 2020.

[57] Angelo Porrello, Luca Bergamini, and Simone Calderara. Robust re-identification by multiple views knowledge distillation. In *ECCN*, pages 93–110. Springer, 2020.

[58] Guanglu Song, Biao Leng, Yu Liu, Congrui Hetang, and Shaofan Cai. Region-based quality estimation network for large-scale person re-identification. In *AAAI*, 2018.

[59] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, pages 1318–1327, 2017.

Zhikang Wang is currently a master with the School of Electronic Engineering at Xidian University. He was a visiting scholar at Learning and Vision Lab, National University of Singapore (NUS) from 2019 to 2020 under the supervision of Dr. Jiashi Feng. He joined the Pensees Singapore as an intern in 2019. His research interests include computer vision, deep learning, and multimedia data processing.

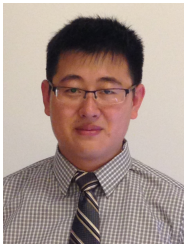
Lihuo He received the B.Sc. degree in electronic and information engineering and the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, China, in 2008 and 2013, respectively. He is currently an Associate Professor with Xidian University. His research interests focus on image/video quality assessment, cognitive computing, and computational vision.

Xiaoguang Tu is currently a lecturer in Aviation Engineering Institute at Civil Aviation Flight University of China. He received his Ph.D degree from the University of Electronic Science and Technology of China (UESTC) in 2020. He was a visiting scholar at Learning and Vision Lab, National University of Singapore (NUS) from 2018 to 2020 under the supervision of Dr. Jiashi Feng. His research interests include convex optimization, computer vision, and deep learning.

Jian Zhao received the Bachelor degree from Beihang University in 2012, the Master degree from the National University of Defense Technology in 2014, and the Ph.D. degree from the National University of Singapore in 2019. He is currently an Assistant Professor with the Institute of North Electronic Equipment, Beijing, China. His main research interests include deep learning, pattern recognition, computer vision, and multimedia analysis. He has published over 40 cutting-edge papers. He has received the Young Talent Support Project from China Association for Science and Technology, and Beijing Young Talent Support Project from Beijing Association for Science and Technology, the Lee Hwee Kuan Award (Gold Award) on PREMIA 2019, the Best Student Paper Award on ACM MM 2018, and the top-3 awards several times on worldwide competitions. He is the EAC of VALSE, and the committee member of CSIG-BVD. He has served as the invited reviewer of NSFC, T-PAMI, IJCV, NeurIPS (one of the top 30% highest-scoring reviewers of NeurIPS 2018), CVPR, etc.

Xinbo Gao received the B.Eng., M.Sc., and Ph.D. degrees in signal and information processing from Xidian University, China, in 1994, 1997, and 1999, respectively. From 1997 to 1998, he was a Research Fellow at the Department of Computer Science, Shizuoka University, Shizuoka, Japan. From 2000 to 2001, he was a Post-doctoral Research Fellow at the Department of Information Engineering, the Chinese University of Hong Kong, Hong Kong. Since 2001, he has been at the School of Electronic Engineering, Xidian University. He is currently a Cheung Kong Professor of Ministry of Education, a Professor of Pattern Recognition and Intelligent System, and the Director of the State Key Laboratory of Integrated Services Networks, China. His current research interests include multimedia analysis, computer vision, pattern recognition, machine learning, and wireless communications. He has published six books and around 200 technical articles in refereed journals and proceedings. Prof. Gao is on the Editorial Boards of several journals, including Signal Processing (Elsevier) and Neurocomputing (Elsevier). He served as the General Chair/Co-Chair, Program Committee Chair/Co-Chair, or PC Member for around 30 major international conferences. He is a Fellow of the Institute of Engineering and Technology and a Fellow of the Chinese Institute of Electronics.

Shengmei Shen received the Graduate and Master degrees from Xidian University, China, in 1986 and 1988, respectively. She is currently working as the chief scientist and managing director at Pensees Singapore, leading AI technology development especially in deep learning for surveillance, automotive, robotics, and other applications with smart innovation and solution.



Jiashi Feng is currently an Assistant Professor in the Department of Electrical and Computer Engineering at the National University of Singapore. He received his Ph.D. from the National University of Singapore in 2014. Before joining NUS as a faculty, he was a postdoc research fellow at UC Berkeley. Dr. Feng's research areas include computer vision and machine learning. In particular, he is interested in object recognition, detection, segmentation, robust learning, and deep learning.