Person Re-Identification by Context-aware Part Attention and Multi-Head Collaborative Learning

Dongming Wu, Mang Ye, Gaojie Lin, Xin Gao, and Jianbing Shen, *Senior Member, IEEE*

***Abstract*—Most existing works solve the video-based person re-identification (re-ID) problem by computing the representation of each frame independently and finally aggregate the frame-level features. However, these methods often suffer from the challenging factors in videos, such as serious occlusion, background clutter and pose variation. To address these issues, we propose a novel multi-level Context-aware Part Attention (CPA) model to learn discriminative and robust local part features. It is featured in two aspects: 1) the context-aware part attention module improves the robustness by globally capturing the relationship among different body parts across different video frames, and 2) the attention module is further extended to multi-level attention mechanism which enhances the discriminability by simultaneously considering low- to high-level features in different convolutional layers. In addition, we propose a novel multi-head collaborative training scheme to improve the performance, which is collaboratively supervised by multiple heads with the same structure but different parameters. It contains two consistency regularization terms, which considers both multi-head and multi-frame consistency to achieve better results. The multi-level CPA model is designed for feature extracting, while the multi-head collaborative training scheme is designed for classifier supervision. They jointly improve our re-ID model from two complementary directions. Extensive experiments demonstrate that the proposed method achieves much better or at least comparable performance compared to the state-of-the-art on four video re-ID datasets.**

***Index Terms*—Person re-identification, multi-level spatial-temporal attention, context-aware part attention.**

I. Introduction

Person re-identification (re-ID) is the task of identifying the same person in different images or videos, taken from different and non-overlapping cameras. It has received increased attention in recent years due to the increasing demand of public safety and rapidly growing surveillance camera networks. In recent years, a large amount of works have been proposed to tackle this problem under image setting and prominent progresses have been achieved. Most of these image-based methods focus on extracting distinctive feature representation from pedestrian images [1], [2], learning robust distance metric for similarity measurement [3], [4] or combining both of them into an end-to-end deep convolutional neural network (CNN) [5], [6], [7], [8], [9].

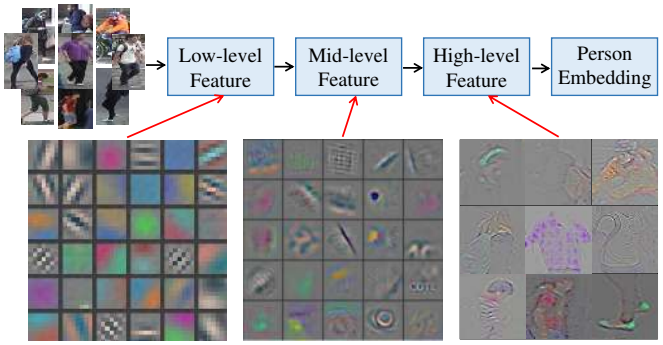

Fig. 1. Visualization of the features learned in different convolutional layers. It is observed that CNN captures low- to high-level features in shallow to deep layers in re-ID tasks.

Despite decades of efforts, person re-ID is still a challenging problem because of the variations in human pose, view angles, background clutters, occlusions and illumination condition. Temporal information is useful to resist and eliminate these complex spatial noise, while it is usually ignored in image-based re-ID methods. With the emergence of large-scale video re-identification datasets [10], [11], several studies [12], [13], [14], [15], [16], [17], [18], [19] focus on solving the person re-ID problem under video setting. Since a video clip usually contains much richer spatial-temporal information, it is beneficial to identify a person under complex environment and serious appearance variations [10]. In addition, person Re-ID is mostly applied in video surveillance, where video is usually the first-hand materials captured by surveillance cameras, so it is more convenient to deal with this problem directly on videos. In this paper, we attempt to tackle the person re-ID task under video setting.

An efficient way to solve person re-ID task on large-scale video-based datasets is to learn a mapping function to convert video-clips into lower dimensional feature space, so that the similarity can be measured by computing the Euclidean Distance between different video embeddings. The widely-used technique is to extract the frame-level appearance features with deep convolutional neural networks and then adopt the average or maximum pooling [20], [11] to aggregate frame-level features and obtain the video representation. However, this strategy is sensitive to outlier noisy frames caused by detection/tracking error and occlusion. To address these issues, some works use attention module [14], [15], [16], [21] or Recurrent Neural Network (RNN) [12], [13] to capture robust spatial-temporal cues across multiple video frames and weaken the influence from corrupted frames ($e.g.$ occlusion). To

D. Wu and G. Lin are with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China.

M. Ye is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong, China. (Email: mangye16@gmail.com)

X. Gao is with Computer, Electrical, and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.

J. Shen is with Inception Institute of Artificial Intelligence, Abu Dhabi, UAE, and also with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing, China. (Email: shenjianbingcg@gmail.com)

further improve the discriminability, some recent works [18], [17] utilize part attention to extract discriminative local features based on different body parts in multiple video frames. However, all these methods only utilize the high-level feature (*the output of last convolutional layer*) to capture the attention cues, while such features are easily being contaminated due to error accumulation in shallow convolutional layers, especially for large-scale video re-ID tasks with a large amount of noisy frames, which results in limited robustness. Moreover, the informative low-level features (*the output of shallow convolutional layers*) [22] are ignored in their attention module, but it has been shown that different layers capture different kinds of discriminative features in Fig. 1. Therefore, it motivates us to investigate a solution to simultaneously capture the robust and discriminative multi-level part attention cues in different layers from both spatial and temporal domains.

In this paper, we propose a novel multi-level context-aware part attention model for robust and discriminative video feature representation learning. Specifically, the Context-aware Part Attention (CPA) module is designed to extract discriminative local part features by globally considering the spatial-temporal context information of the local body parts across different video frames. The CPA module captures the spatial-temporal context information via fully exploiting the relationship among multiple local parts across different frames. After that, the learned contextual knowledge is aggregated to local activation to facilitate local feature learning. To utilize the informative multi-level features, we further extend the CPA module into a multi-level attention mechanism by plugging it into different convolutional layers of the feature network in a residual connection manner, which considers simultaneously the discriminative cues in different feature levels. In this way, our multi-level CPA framework can also resist the spatial noisy in an early stage during feature extracting process.

While existing methods capture the most salient features referring to the training benchmark with one classifier head, they miss lots of useful and local regions. To overcome this limitation, we expect to mine these new features with different classifier heads focusing on different regions. For that, we propose a multi-head collaborative learning scheme to enhance the classifier supervision and improve the generalization ability of the feature network during training. The basic idea is to adopt multiple heads with the same structure but different initialization parameters to collaboratively optimize the network parameters. Incorporated with multi-head consistency regularization, each head can transfer their knowledge to each other, which collaboratively improves the accuracy but without extra model architecture design. It also contains a multi-frame consistency regularization to eliminate the effect of outlier frames in a video sequence, which improves the robustness to noisy frames. Feature extracting and classifier supervision are two most important parts for training a re-ID model. The multi-level CPA is designed for feature extracting by improving the backbone's capacity to model abundant spatial-temporal information, while the multi-head training scheme is designed for classifier supervision by incorporating multiple supervision heads to improve the performance in a collaborative learning manner. Both of them jointly improve our re-ID model from two complementary directions.

In summary, our method brings following contributions:

- We propose a novel context-aware part attention (CPA) module to extract robust and discriminative part features from multiple video frames by considering the context information in both spatial and temporal domains, which can be seamlessly plugged into low-level convolutional layers with a residual mechanism.
- We further boost the CPA module to multi-level by exploiting the informative features in different convolutional layers, which results in much better performance than existing single layer counterparts.
- We propose a new multi-head collaborative learning scheme to improve the accuracy without extra model architecture design. It contains two novel consistency constraints, which enforces the network to utilize the multi-head collaboration and multi-frame information.
- We demonstrate the effectiveness of our approach on four challenging video re-identification datasets. The proposed method outperforms all the other state-of-the-art methods under multiple evaluation metrics.

II. Related Work

A. Image-based Person Re-identification.

Person re-ID based on still image has been extensively explored in the literature [23], [24], [25], [26], [27], [6], [28], [29], [30] during the past few years. Traditional methods for image-based re-ID can generally be categorized into two parts: discriminative feature learning [31], [32], [1], [33], [27] and robust metric learning [6], [28], [29], [30], [34]. In the part of discriminative feature learning, researchers focus on designing hand-crafted appearance feature representation [1], [35], [36], [37], such as color histograms, local binary patterns and Gabor features. which are robust to illumination and viewpoint changes. Besides, the methods of robust metric learning aim to learn a discriminant subspace or an integrated metric to encourage the distance of pedestrian from the same identity be small and those from different identities be large [38], [4], [29], [3]. With the emergence of deep learning, convolutional neural networks have been applied to person re-ID to learn discriminative feature representation and robust distance metric jointly, which has made great progress. Contrast loss [5], [39], triplet loss [40], [7], [6] and softmax loss [41], [42], [43] are available to be used as supervision and help networks learn robust feature by end-to-end training. Moreover, recent works integrate part-based features, which come from uniform partions of image or feature map, with global feature to form the final feature representation [44], [45]. To lower the requirement of manually annotated data, recent unsupervised approaches are investigated through the strategies of style transfer and clustering [46].

B. Video-based Person Re-identification.

Recently, more and more works pay attention to the video-based person re-ID [47], [48], [49], [50], [51], which is an extension of the image-based one. Existing studies [13], [14],

[20], [12], [15], [16], [19] adopt convolutional neural network (CNN) to extract the appearance representation of each frame and mainly focus on different techniques to fuse temporal information. For examples, McLaughlin *et al.* [13] adopt Recurrent Neural Network (RNN) to pass the message among frame-level features and utilize temporal maximum pooling to obtain video descriptors. Liu *et al.* [14] propose a Quality Aware Network (QAN) to estimate quality score of each frame automatically and use different weights to fuse frame-level features. Chen *et al.* [19] introduce a temporal co-attention which weights frame features by considering the mutual influence between gallery and query sequences. Hou *et al.* [52] try to mine all the salient parts of one video clip by erasing the activation parts of previous frames. However, all these methods focus only on designing temporal model to fuse frame-level features, and lack the capabilities of discovering discriminative local body parts in each frame. Although recent works consider these local parts with hierarchy architecture [53] or graph model [54], they cannot be deployed flexibly like our method.

C. Attention Models for Video Re-ID.

Attention model is a very popular technique to exploit discriminative local feature in deep neural network [55], [56], [57], [58]. Recent works [18], [17], [59], [60] in video-based person re-ID incorporate spatial-temporal attention to take advantage of discriminative local features to learn robust video embeddings. Li *et al.* [18] propose a diversity regularized spatial attention model to extract appearance representation of different parts and then use temporal attention to fuse part-level representation. Fu *et al.* [17] propose a parameter-free spatial-temporal attention (STA) module, which assigns attention score for each spatial region to achieve discriminative parts mining and frame selection. Zhang *et al.* [61] presents a non-parametric attention mechanism for video re-ID with a generalized pairwise similarity measurement by the binary classifier. All these works exploit only the spatial-temporal information in high-level features, ignoring the abundant spatial-temporal cues in different semantic levels. Liao *et al.* [59] use 3D CNN to extract the aggregate representation of spatial-temporal features and incorporate block as spatial-temporal attention strategy. However, the stacked 3D CNN layers result in substantial growth of parameters which not only make their model computationally expensive, but also leads to the difficulty in model training and optimization. Also, the non-local block aims to model the long-range dependency in pixel level, which further increases the demand of computational resources. As a result, their model is in demand of training on eight GTX TITAN X GPUs and obtains an inferior performance. To mitigate the shortcomings of 3D CNN, Li *et al.* [60] propose a Multi-scale 3D (M3D) convolution layer as a more efficient alternative to model the temporal cues for each local activation. Additionally, a Residual Attention Layer (RAL) is proposed to jointly learn spatial and temporal attention masks to further refine the learned temporal cues. However, their model ignores the relationship of different local features which is beneficial to learn the context information.

Different from these works, our Context-aware Part Attention (CPA) module exploits relationship among different body parts and aggregates the contextual information in both spatial and temporal domains, which is quite suitable to learn part discriminative features for video re-id. We incorporate CPA to leaned spatial-temporal attention at multi-levels to extract discriminative part features at different convolutional layers rather than a single pooling layer. Experiments shown in Table III, multi-level attention strategy consistently improves the re-id performance in all the settings. Also, since the number of body parts is far less than the number of pixels, our CPA module considers spatial-temporal context information of different feature levels in a more efficient way, which makes it more easy and flexible to be incorporated.

D. Collaborative Learning.

An ensemble of multiple instances of a target neural network trained with different random seeds generally yields better predictions than a single trained instance. However, it is also very computationally expensive. To lower computational complexity, several training techniques have been developed by adding additional networks only in the training graph to boost accuracy, including auxiliary training [62] and knowledge distillation [63]. In [64], Song *et al.* propose a collaborative learning scheme, which provides a simple but effective multi-head training framework for any given architecture to improve accuracy. The multi-head structure gives a strong supervision to the shared backbone, and it is demonstrated robust to noise labels. Inspired by this work, we propose to use a multi-head collaborative training scheme to improve the performance of video re-ID, which is supervised by multiple heads with the same structure but different parameters. It also contains a multi-head consistency loss and a multi-frame consistency loss to regularize the ambiguity among multi-frame feature learning and multi-head identity prediction.

III. Our approach

The proposed video-based person re-ID method mainly contains two parts: multi-level context-aware part attention module (Sec. III-A) and multi-head collaborative learning scheme (Sec. III-B). The former introduces a backbone network with multi-level context-aware part attention (CPA) module to enhance the feature extracting, where CPA captures robust and discriminative part features by utilizing the context information in both spatial and temporal domains The latter presents our training scheme with multi-head collaborative learning framework and two kinds of regularization strategies to enhance the classifier supervision and improve the generalization ability of our re-ID model.

A. Multi-level Context-aware Part Attention

We firstly introduce the proposed Context-aware Part Attention (CPA) module, which aims at extracting robust and discriminative local part features by utilizing the context information in both spatial and temporal domains. Given an input video sequence with T frames, we feed it into residual blocks and obtain a set of frame-level feature maps. We decompose the feature map of each person image into M non-overlapping

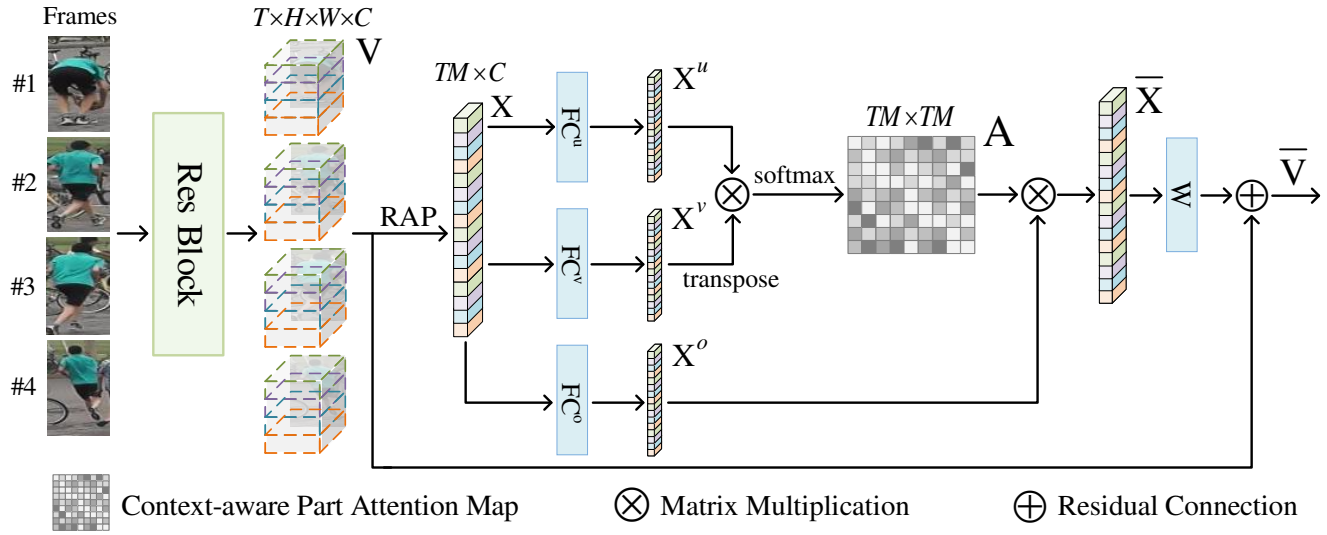

Fig. 2. Illustration of Context-aware Part Attention (CPA) module. Given the intermediate feature map V of a person tracklet in a specific layer, we first divide it into M non-overlapping and equal-size regions vertically, then apply region-based average pooling (RAP) and reshaping to obtain part-level feature map X. X^u, X^v and X^o are generated by feeding X through three different linear transformations. The context-aware part attention between local parts can be computed efficiently by matrix multiplication between X^u and the transpose of X^v followed by a row-wise softmax function. By using learned context-aware part attention as weights to linearly aggregate the part-level representations in X^o, we obtain the refined part features which contain global context knowledge and encode part relationship. The final part features output from the linear transformation W are reshaped, up-sampled and added back to original input V through a residual connection.

parts. CPA aims at learning a context-aware attention map with shape of $TM \times TM$ for each local parts cross spatial and temporal domains. The structure of CPA module is illustrated in Fig. 2.

The original frame-level feature maps are denoted by $V = \{v_t | v_t \in \mathbb{R}^{h \times w \times C}\}_{t=1}^{T}$, where $h \times w$ is the resolution of the feature map, C is the channel dimension and v^t can be viewed as a 3D tensor of activation. Notice that since pedestrian in images can be decomposed into several body parts from head to foot, we simply slice each v_t into M equal-size and non-overlapping regions vertically. A region-based average pooling (RAP) is applied to each region to obtain a set of part-level features $X = \{x_i | x_i \in \mathbb{R}^C\}_{i=1}^{TM}$, which summarizes the feature statistics of each local part. X is then fed into three different linear layers FC^u, FC^v and FC^o to generate three new part-level feature embeddings.

$$\begin{aligned} X^u &= \{x_i^u | x_i^u \in \mathbb{R}^{\frac{C}{r}}\}_{i=1}^{TM} \\ X^v &= \{x_i^v | x_i^v \in \mathbb{R}^{\frac{C}{r}}\}_{i=1}^{TM} \\ X^o &= \{x_i^o | x_i^o \in \mathbb{R}^{\frac{C}{r}}\}_{i=1}^{TM} \end{aligned} \tag{1}$$

Note that three linear layers share the same structure but with different parameters. And they all reduce the channel dimensionality of X with ratio r. X^u and X^v are used to calculate the context-aware part attention $A = \{a_{i,j}\}_{i=1,j=1}^{TM}$ for each pair of part feature, and X^o serves as a contextual representation to be transferred to other local parts.

Specifically, the context-aware part attention between part i and part j is computed by the inner-product between x_i^u and x_j^v.

$$a_{i,j} = \frac{\exp((x_i^u)^T(x_j^v))}{\sum_{k=1}^{T} M \exp((x_i^u)^T(x_k^v))} \tag{2}$$

where softmax function is further used to normalize the value of attention for each body part. Compared to the ℓ_2 normalization in STA [17], the softmax normalization enlarges the attention discrepancy between discriminative parts, which has better discriminability. The context-aware part attention A can also be computed more efficiently by performing a matrix multiplication between X^u and the transpose of X^v followed a row-wise softmax. Each row of the context-aware attention map $A = \{a_{i,j}\}_{i=1,j=1}^{TM}$ indicates the attention from part i to every other part j. If we calculate A through inner product of part-level feature from the same branch, then the value of part attention $a_{i,j}$ will be the same as that of $a_{j,i}$. However, the value of part attention $a_{i,j}$ and $a_{j,i}$ should be different. For example, the noisy patch should pay more attention to the informative patch to get contextual information for feature reparation, while the informative patch does not need to pay attention to the noisy patch. Therefore, we adopt part-level feature embeddings, X^u and X^v, from two different branches rather than the same branch to compute the context-aware attention map A.

With the calculated context-aware part attention, the part features are updated by the linear combination of the contextual representation from different parts

$$\bar{x}_i = \sum_{j=1}^{N} a_{i,j} x_j^o \tag{3}$$

where $\bar{x}_i$ is the weighted summation of contextual representation x_j^o over all parts guided by the context-aware part attention. Therefore, the refined part features is able to consider the global context information and the relationship between different part features across both spatial and temporal domains. The "context-aware" in the name of CPA means the relationship among multiple local parts across different frames rather than the background information. Note that this process can be efficiently computed by matrix multiplication between A and the transpose of X^o without much computation cost.

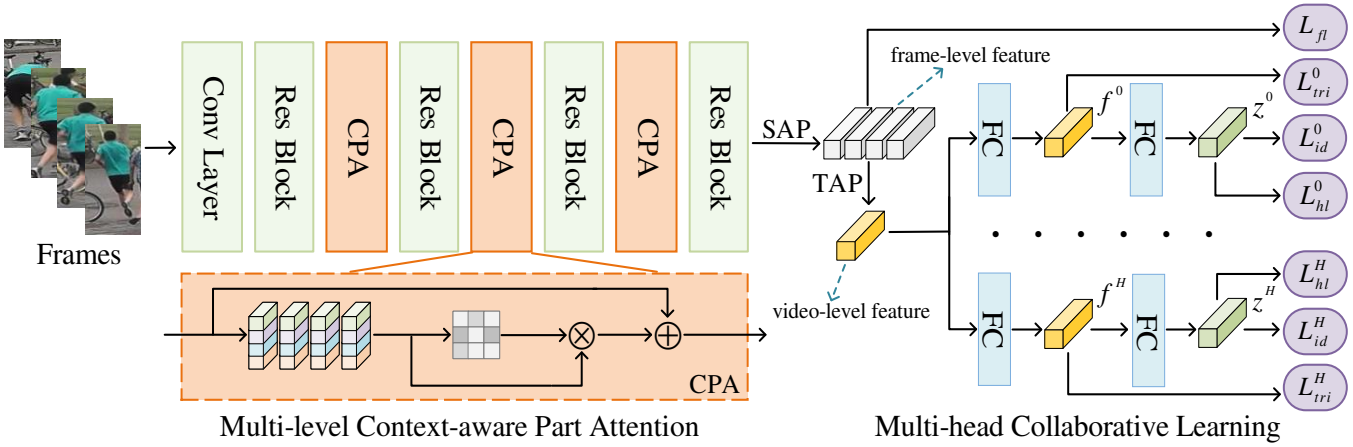

Fig. 3. The overview of our approach. It is mainly comprised of two parts: multi-level context-aware part attention feature network and multi-head collaborative learning scheme. The CPA module is seamlessly plugged into different stages of the backbone network to learn multi-level context-aware part attention. SAP represents the spatial average pooling to get the frame-level feature vectors and TAP represents the temporal average pooling to get the video-level feature vector. Several supervision heads are applied to the video-level feature simultaneously to provide more robust supervision. Each learning head consists of two fully connected layers named embedding layer and classification layer respectively. During training, each head is supervised by an identity classification loss, hard triplet loss and multi-head consistency loss respectively. Besides, multi-frame consistency loss is used to regularize the frame-level feature after SAP. During testing, the feature vectors after the embedding layer from all heads are concatenated together as the final descriptor of the input video. The multi-head framework only leads to minor computational growth during inference as shown in Table VIII.

After we get the re-weighted part features $\bar{X}$, we feed it into a learnable linear transformation $W \in \mathbb{R}^{C \times \frac{C}{r}}$. To seamlessly connect with different residual blocks, we introduce a residual connection to combine the attention refined part feature maps and the original feature maps V, which is denoted by

$$\bar{V} = V + \tau(\bar{X}W) \tag{4}$$

where $\tau(\cdot)$ denotes the reshape and resize operator to make sure the resolution being consistent with the input feature maps. The contextual knowledge of part relationship is brought back to the original feature map for representation enhancement. The final local representations of different body parts achieve mutual gains and are more robust to occlusion and pose variant.

After all these operations, part-level features have a global contextual effect and selectively aggregate the part features across spatial and temporal domains according to the context-aware part attention map. To capture the multi-level attention cues in different feature levels, we extend the CPA module into a muti-level attention mechanism by plugging the CPA module into different stages of the convolutional neural network. In particular, ResNet-50 has one convolutional block ($conv1$) and four residual blocks ($conv2\_x, 3\_x, 4\_x, 5\_x$), and we plug three CPA modules after $conv2\_x, 3\_x, 4\_x$ respectively. In addition, other convolutional neural networks are also feasible to be adopted as backbone network.

Different from the the Non-Local structure [55], which is designed to capture long-range dependencies between distant pixels, while our CPA module aims to exploit the relationship between different body parts. The dependencies between pixel-level feature is too sensitive to learn in video re-id model, because the serious background clutter in pedestrian video can dominate and contaminate the relationship learning process. While the part-level feature utilizes the prior knowledge of body structure to summarize the appearance characteristic of each body part and reduce the effect of background clutter. Therefore, our scheme shows better robustness against the background clutter than the Non-Local structure. In addition, the efficiency is also greatly improved. The number of part-level feature ($T \times M \times 1$) is much less than the number of pixel-level feature ($T \times h \times w$), which makes the relationship easier to be learnt. Also, the time complexity and space complexity of our CPA module have been reduced from $\mathcal{O}(T^2 \times h^2 \times w^2)$ to $\mathcal{O}(T^2 \times M^2)$. As shown in Table VIII, the CPA module performs significant better than Non-Local block with much less FLOPS and the required GPU memory. Our CPA module takes a more efficient way to exploit the spatial and temporal context information, which makes it more computationally efficient for video re-ID task.

B. Multi-head Collaborative Learning

We now introduce the training scheme with Multi-head Collaborative learning (MHC). In Fig. 3, with the output video-level feature representation of temporal average pooling (TAP) layer, we propose to use multiple supervision heads rather than single head to guide the feature learning process. The main motivation is that training with single head (classification layer) is easy to overfit to local minima, and the supervision information from classification layer to backbone is limited. MHC optimizes the network with randomly initialized multiple classifier layers to avoid overconfident learning on a single head.

In our multi-head collaborative learning framework, each head has the same design and supervision but with different parameters. Each head is consist of two fully connected layers named embedding layer and classification layer respectively. A Batch Norm, ReLu and Dropout layer are inserted before classification layer. For the sake of brevity, we don't draw them out in Fig. 3. The multi-head collaborative learning scheme enables diversity predictions on the same sample, and each head may focus on different patterns to identify each

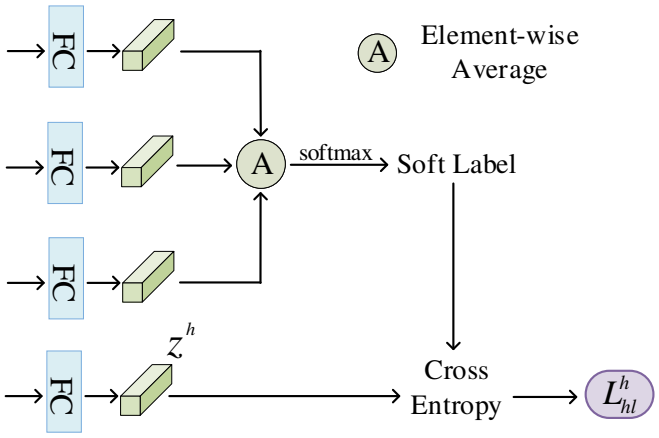

Fig. 4. Illustration of multi-head consistency loss. The prediction logits z from all other learning heads are averaged to obtained a soft label, which presents the prediction consensus of other heads. The cross entropy between z^h and the soft label are calculated as multi-head consistency loss.

sample, which provides stronger supervision and guides the backbone feature network to learn more robust and general deep features. Thus, the overfitting cost caused by the single identity classification layer are reduced. During testing, the feature vectors after the embedding layer from all heads are concatenated together as the final descriptor of the input video.

The learning objective of multi-head collaborative learning contains two main parts: 1) baseline loss: triplet loss (L_{tri}) with hard mining [6] and identity classification loss (L_{id}) with softmax cross-entropy; and 2) consistency regularization: multiple-head consistency loss (L_{hl}) and multiple-frame consistency loss (L_{fl}).

1) Baseline Loss: **Triplet loss.** The triplet loss with online hard mining aims at capturing the relationship between different video sequences [6]. Specifically, we randomly sample P identities and for each identity randomly sample K tracklets to form a batch of N_b samples, where $N_b = P \times K$. Typically, the loss function for each head h is formulated as follows

$$L^h_{tri} = \sum_{i=1}^{P}\sum_{a=1}^{K}\Big[\alpha + \overbrace{\max_{p=1\cdots K} ||f^h_{i,a} - f^h_{i,p}||_2}^{hardest\ positive} - \underbrace{\min_{\substack{n=1\cdots K \\ j=1\cdots P \\ j\neq i}} ||f^h_{i,a} - f^h_{j,n}||_2}_{hardest\ negative}\Big]_+ \tag{5}$$

where $[\cdot]_+ = \max(\cdot, 0)$, α is the margin between positive and negative pairs, $f^h_{i,a}$, $f^h_{i,p}$ and $f^h_{i,n}$ are the feature embeddings of the anchor, positive and negative samples respectively. We use positive and negative to denote the samples with same or different identities from the anchor samples.

Identity loss. The identity loss aims at capturing the identity invariant component in the learned features. The embedding feature f^h is fed into the classification layer to obtain classification prediction logit z^h, which is supervised by an identity classification loss. The identity loss for each head h is denoted by

$$L^h_{id} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{k=1}^{N_{id}} y_{i,k} \log \frac{\exp(z^h_{i,k})}{\sum_{l=1}^{N_{id}} \exp(z^h_{i,l})} \tag{6}$$

where y_i is the one-hot ground truth label of sample i in a batch, z^h_i is the prediction logit and N_{id} is the total category number of person identities.

2) Consistency Regularization: **Multi-head consistency loss.** The multi-head consistency provides regularization to each head, which aims at collaboratively improving the feature learning in multiple heads. The pipeline of multi-head consistency loss is illustrated in Fig. 4. Since the multi-head training framework has multiple prediction results, these results can be averaged to obtain a soft label which presents the consensus of multiple classifier heads. We use the soft label to supervise each classifier head as multi-head consistency loss, encouraging each classifier head to learn from the multi-head prediction consensus. Specifically, the soft label $\bar{y}^h$ is computed by averaging the identity predictions z^h over all other heads followed by a softmax non-linearity. $\bar{z}^h$ and $\bar{y}^h$ are defined by

$$\bar{z}^h_k = \frac{1}{H-1}\sum_{j\neq h} z^j_k \tag{7}$$

$$\bar{y}^h_k = \frac{\exp(\bar{z}^h_k)}{\sum_{l=1}^{N_{id}} \exp(\bar{z}^h_l)} \tag{8}$$

where H is the total number of learning heads.

The proposed multi-head consistency regularization measures the distance between multi-head prediction consensus and single-head prediction, which aims at transferring the learned information among multiple heads. Similar to the identity classification loss, we use the soft label as the supervision by computing the softmax cross entropy between soft label $\bar{y}^h_i$ and the identity prediction z^h_i. Our multi-head consistency loss is defined by

$$L^h_{hl} = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{k=1}^{N_{id}} \bar{y}^h_{i,k} \log \frac{\exp(z^h_{i,k})}{\sum_{l=1}^{N_{id}} \exp(z^h_{i,l})} \tag{9}$$

During back-propagation, the gradients from multiple heads are gathered into the backbone extractor to provide more reliable gradient information and facilitate the extractor to learn more robust global feature. In the inference stage, we concatenate the f^h from all the heads together as video feature descriptors.

Multi-frame consistency loss. The multi-frame consistency provides regularization on the intuition that the feature representations of frames from the same video sequence should be similar to each other, and can therefore help enhance the robustness against outlier frames. We compute the Euclidean distance between the frame-level features as multi-frame consistency loss to restrict the differences between frames. Then, the backbone feature network can focus on the common patterns across frames rather than the noisy patches during feature extracting process. The multi-frame consistency loss for an input video sequence k is denoted by

$$L^k_{fl} = \sum_{t=1}^{T}\sum_{s=1}^{T} \|\ell_2(f_t) - \ell_2(f_s)\|_2 \tag{10}$$

where f_t is the features extracted from the t^{th} frame of a person tracklet after spatial global average pooling as depicted in Fig. 3. $\ell_2(\cdot)$ denotes ℓ_2 normalization.

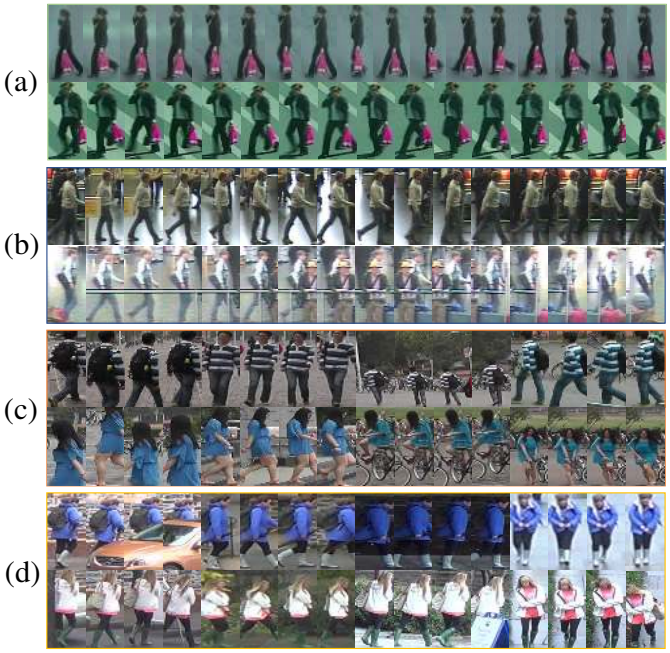

Fig. 5. Illustration of video frames sampled from tracklets in four video person re-ID datasets: (a) PRID2011; (b) iLIDS-VID; (c) MARS; (d) Duke-VideoReID.

The multi-frame consistency loss forces backbone extractor to fully utilize the context information learned from CPA modules and pay more attention to consistent pattern among the person video. As a result, the ambiguity among learned frame-level features caused by some frames' corrupted regions is relieved. And the final representation is more discriminative and robust to occlusion, blur, illumination and pose changes.

3) Total Loss: In our method, the identity loss and triplet loss are used as baseline loss for jointly representation learning and metric learning, while the two kinds of consistency regularization loss are specifically designed for our multi-head video re-ID training framework. The multi-frame consistency loss regularizes the feature extracted from different frames, while the multi-head consistency loss regularizes the identity prediction from different classifier heads. Both of them are combined with two baseline loss functions to jointly train the whole framework. The total loss L for our multi-head collaborative learning objective framework is defined by

$$L = \sum_{h=1}^{H}(L_{tri}^{h} + L_{id}^{h} + \beta L_{hl}^{h}) + \sum_{k=1}^{K}\gamma L_{fl}^{k} \qquad (11)$$

where β and γ are coefficients to adjust the contribution of L_{hl}^{h} and L_{fl} to the total loss respectively. K represents the total number of video sequences in each batch.

IV. Experimental Results

In this section, we thoroughly analyze the effectiveness our method on four challenging video person re-ID dataset including PRID2011, iLIDS-VID, MARS and Duke-VideoReID. Firstly, to validate the superiority of our method, we compare our approach with other state-of-the-art video re-ID methods on four challenging video re-ID datasets. We then conduct extensive ablation experiments to demonstrate the effectiveness of each component of our proposed approach. In addition, we analyze the effect of different parameter setting in our approach, visualize the learned part attention map and analyze the computation efficiency of our framework. At last, we also evaluate the generalization ability of our approach in cross-dataset experiments.

A. Datasets and Evaluation Protocol

The PRID2011 dataset [65] is a relatively old and small video-based person re-identification dataset. It is collected in outdoor scenes with relatively simple backgrounds but large illumination and viewpoint change, as shown in Fig. 5 (a). The pedestrian tracklets of this dataset are captured by two non-overlapping surveillance cameras. The camera A captures 385 identities and the camera B captures 749 identities. There are 400 tracklets of 200 pedestrians captured by both camera A and B. Each pedestrian sequence is consisting of 5 to 675 frames with an average number of 100. we only select the 178 identities of which the length of pedestrian tracklet is more than 27 frames.

The iLIDS-VID dataset [66] is also a relatively old and small video-based person re-identification dataset which consists 600 tracklets of 300 distinct individuals under two non-overlapping cameras, with bounding boxes annotated by humans. Each tracklets has 23 to 192 frames and the average number is 73. Some samples of iLIDS-VID are shown in Fig. 5 (b). Following the standard evaluation protocol of iLIDS-VID and PRID2011, the video clips captured by the first camera are regard as probe set and those captured by the other are regard as gallery set. All the identities in dataset are randomly split into 50% for training and 50% for testing. We repeat the procedure for 10 times with different test/train splits and obtain the final average results.

The MARS dataset [10] is one of the largest video-based person re-identification datasets. It consists of 1,261 pedestrians and 20,751 tracklets captured by six non-overlapping cameras on Tsinghua campus, as shown in Fig. 5 (c). Each identity is captured by at least 2 cameras and has 13.2 sequences on average. The bounding boxes in MARS dataset are generated automatically by DPM detector and GMMCP tracker. 3,248 tracklets are used as distractors due to the failure of detection or tracking. The 1,261 pedestrians are split into 625 and 636 identities for training and testing.

The Duke-VideoReID dataset [11] is a newly released large-scale video-based person re-identification dataset derived from the Duke dataset [67]. Some samples are shown in Fig. 5 (d). It consists of 4,832 tracklets from 1,812 identities and each identity only has one tracklet under a camera, with bounding boxes annotated manually. It is split into 702 identities for training, 702 identities for testing and 408 identities as the distractors. Totally, there are 2,196 tracklets for training, and 2,636 tracklets for testing and distractors.

Evaluation Metrics. We employ the Cumulative Matching Characteristic curve (CMC) and the mean Average Precision score (mAP) as evaluation criteria. CMC considers re-ID as a ranking problem and represents the accuracy of the person retrieval with each given query. Since the tracklets of both

TABLE I
PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE PRID-2011 AND ILIDS-VID DATASET. DS REPRESENTS THE SENSE SAMPLING STRATEGY DURING EVALUATION STAGE. THE RANK-1, -5, -10, -20 ACCURACY SCORES (%) ARE REPORTED. OF IS SHORT FOR OPTICAL FLOW.

Classes	Method	PRID-2011				iLIDS-VID			
		R1	R5	R10	R20	R1	R5	R10	R20
Traditional	DVDL [34]	40.6	69.7	77.8	85.6	25.9	48.2	57.3	68.9
	DVR [66]	48.3	79.4	87.3	94.4	41.3	63.5	72.7	83.1
	TDL [20]	58.6	80.8	87.4	93.3	56.2	88.2	95.3	97.8
	STFV3D+KISSME [68]	62.5	83.6	88.1	89.9	44.3	71.7	83.7	91.7
	RFANet+RSVM [12]	58.2	85.8	93.4	97.9	49.3	76.8	85.3	90.0
	LMKDCCA [69]	86.4	97.5	99.6	100	73.3	90.5	94.7	98.1
DL-based	CNN+Kiss.+MQ [10]	70.0	81.4	-	95.1	53.0	81.4	-	95.1
	CNN-RNN[13]	70.0	90.0	95.0	97.0	58.0	84.0	91.0	96.0
	SeeForest [15]	79.4	94.4	-	99.3	55.2	86.5	-	97.0
	ASTPN [16]	77.0	95.0	99.0	99.0	62.0	86.0	94.0	98.0
	QAN [14]	90.3	98.2	99.3	99.8	68.0	86.8	95.4	97.4
	RQEN [70]	91.8	98.4	99.3	99.8	77.1	93.2	97.7	99.4
	CSACSE [19]	88.6	99.1	-	-	79.8	91.8	-	-
	CSACSE [19]+OF†	**93.0**	**99.3**	**100**	**100**	85.4	96.7	98.8	99.5
	Ours	91.8	98.9	**100**	**100**	85.2	96.8	98.8	**99.8**
	Ours+DS	92.2	99.1	**100**	**100**	**85.8**	**97.1**	**98.9**	**99.8**

† It achieves slightly better performance by adding the optical flow information, which is quite time-consuming and unsuitable for real applications.

TABLE II
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON MARS AND DUKE-VIDEOREID DATASETS. DS REPRESENTS THE SENSE SAMPLING STRATEGY DURING EVALUATION STAGE. THE MAP (%) AND RANK-1, -5, -20 ACCURACIES (%) ARE REPORTED. OF IS SHORT FOR OPTICAL FLOW.

Method	MARS			
	mAP	R1	R5	R20
CNN+Kiss.+MQ [10]	49.3	68.3	82.6	89.4
SeeForest [15]	50.7	70.6	90.0	97.6
QAN [14]	51.7	73.7	84.9	91.6
Latent Parts [71]	56.1	71.8	86.6	93.0
DuATM [72]	62.3	78.7	90.9	95.8
TriNet [6]	67.7	79.8	91.4	-
K-reciprocal [73]	68.5	73.9	-	-
RQEN [70]	71.1	77.8	88.8	94.3
STAN [18]	65.8	82.3	-	-
CSACSE [19]	69.4	81.2	92.1	-
Part-Aligned [31]	75.9	84.7	94.4	97.5
CSACSE [19]+OF	76.1	86.3	94.7	98.2
STA [17]	80.8	86.3	95.7	98.1
NVAN [53]	82.8	**90.0**	-	-
Ours	83.4	88.2	96.4	98.5
Ours+DS	**84.1**	88.2	**96.6**	**98.5**
Method	Duke-VideoReID			
	mAP	R1	R5	R20
ETAP(supervised) [11]	78.3	83.6	94.6	97.6
STA [17]	94.9	96.2	**99.3**	99.6
NVAN [53]	94.9	96.3	-	-
Ours	94.7	96.0	98.9	99.7
Ours+DS	**95.8**	**96.6**	99.1	**99.7**

MARS dataset and Duke-VideoReID dataset are captured from multiple cameras, there may be multiple correct matching results in the ranking list. In this case, mAP is a more suitable and robust metric. The mAP considers person re-ID task as a retrieval problem. We first calculate the average precision (AP) for each query. Then, the mean value of APs of all queries is calculated as the mAP.

B. Implementation details

Sampling strategy. During training, to model more abundant spatial-temporal information and reduce the GPU resource occupy at the same time, we apply the restricted random sampling strategy [18] to generate a compact summary of the pedestrian video sequence. Concretely, given an input video, we first divide it into several segments of equal length, then we randomly sample a frame from each segment to form our input video sequences. In this case, the input video clips enable our model to utilize visual information from the entire video and avoid the redundancy between sequential and neighboring frames. During testing, we use two sampling strategies: in the **First Sampling (FS)** strategy, only the first frame from each segment is used as the test sample to form the input video sequence; in the **Dense Sampling (DS)** strategy, the i^{th} frame from each segment is used as the test sample to form the i^{th} input video sequence, and we average the embedding of all the inputs video sequence to obtain the final video descriptor. DS strategy consider full-scale spatial-temporal information and result more robust video descriptor. However, it's less efficient than FS. For the sake of efficiency and simplicity, we use FS as default sampling strategy unless stated in our experiments.

Network parameter settings. All of our experiments are implemented on the PyTorch platform on Linux with NVIDIA 1080Ti GPU. ResNet-50 pretrained on ImageNet is employed as our baseline model and backbone network. Three CPA modules are further inserted after $conv2\_x$, $conv3\_x$, $conv4\_x$ respectively. The spatial size of each video frame is set as 256×128 pixels in both training and testing stages. The number of spatial regions of each frame in CPA module is set to $M = 4$, and the length of input video sequence is set to $T = 4$ as well. The dimensionality reduction ratio of CPA module is set to $r = 2$. The number of the collaborative learning heads H is set to 6. We recommend to set the margin constant α of the hard-mining triplet loss in Eqn. 5 to 0.3.

The coefficients β and γ are set as 0.3 and 5.0 respectively in Eqn. 11 based on their magnitudes. Dropout rate in each head is set to 0.5.

Training strategy. We randomly sample 8 identities with 4 video sequences to form a mini-batch with batch size 32. Random cropping and random horizontal flipping are used as data augmentation. Stochastic Gradient Descent (SGD) with momentum 0.9 is adopted as the optimizer to train the model for totally 300 epochs. The weight decay factor for L2 regularization is set to 0.0005. We set the initial learning rate as 0.01 and decrease it by 10 times at 100 epochs and 200 epochs respectively. The parameters in BatchNorm layers are also updated in the training phase. The consensus loss is added to the total loss only after 200 epochs, which ensures each head to produce reliable identity predictions.

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

We compare with the state-of-the-art methods on two large-scale video re-ID dataset, MARS and Duke-VideoReID, and two datasets, PRID2011 and iLIDS-VID, as shown in Table II and Table I. Noted that our results are not refined by any post-processing techniques such as re-ranking. And we only take pedestrian image sequences as input, no other extra information like optical flow is utilized. On each dataset, our approach outperforms all the previous state-of-the-arts on both mAP and Rank-1 accuracy. The results suggest that our approach is very effective for video-based person re-identification in challenging scenarios.

Results on MARS. We report the comparison of our approach with twelve state-of-the-art methods, including CNN+Kissme+MQ [10], SeeForest [15], QAN [14], Latent Parts [71], DuATM [72], TriNet [6], K-reciprocal [73], RQEN [70], STAN [18], Part-Aligned [31], CSACSE [19] and STA [17]. As shown in Table II, our approach achieves **84.1%** in mAP and **88.2%** in Rank-1 accuracy, which obtains **3.3%** and **1.9%** improvement in terms of mAP and Rank-1 accuracy respectively compared to previous best results. The prominent improvement demonstrates the effectiveness and superiority of our proposed framework for video person re-ID on large-scale dataset. It also indicates that the multi-level context-aware part attention mechanism and the multi-head collaborative learning are superior to deal with video person re-ID dataset captured under complex environment, such as large gallery set and seriously occlusion.

Results on Duke-VideoReID. The comparison of video re-ID performance on Duke-VideoReID is shown in Table II. Duke-VideoReID is newly released, and there are only two published works, ETAP [11] and STA [17], evaluated on this dataset. Our approach achieves **95.8%** in mAP and **96.6%** in Rank-1 accuracy, which outperforms both of previous two works. By exploiting spatial and temporal information in multi-level of convolution neural network rather than utilizing spatial-temporal attention only after last convolution layer in STA [17], our approach achieves superior performance, with 0.9% improvement in mAP and 0.4% improvement in Rank-1 accuracy respectively.

Results on PRID2011 and iLIDS-VID. On the PRID2011 and iLIDS-VID dataset, we compare our proposed method with fourteen existing state-of-the-art video-based person re-ID methods, including DVDL [34], DVR [66], TDL [20], STFV3D+KISSME [68], RFANet+RSVM [12], LMKDC-CA [69], CNN+Kiss.+MQ [10], CNN-RNN [13], SeeForest [15], ASTPN [16], QAN [14], RQEN [70], STAN [18] and CSACSE [19]. Among all these approaches, the first six methods are traditional methods which uses hand-crafted features, while the others are deep learning based methods and adopt convolutional network to extract frame-level feature. Table I shows the comparison results. Our approach achieves competitive results compared to the most recent work on Rank-1, -5, -10 and -20 accuracy scores on both PRID2011 and iLIDS-VID dataset. Notice that CSACSE [19]+OF achieves a very high performance by incorporating optical flow as extra information to capture the motion information. However, it brings more computation and makes the whole network unable to be trained end-to-end. Comparing to CSACSE [19] without incorporating optical flow, our approach improves the Rank-1 accuracy by **3.6%** and **6.0%** on PRID2011 and iLIDS-VID respectively. Even without utilizing any extra information, our approach still outperforms all previous methods and sets a new state-of-the-art on iLIDS-VID, and achieves state-of-the-art performance on PRID2011. It should be noted that both PRID2011 and iLIDS-VID are relatively old and small, the deep model are overfitting to these two dataset and the performance are almost saturated, especially on PRID2011.

TABLE III
EVALUATION OF THE MULTI-LEVEL CONTEXT-AWARE PART ATTENTION MODULE COMPARING TO THE SINGLE-LAYER ATTENTION MECHANISM AND OTHER ATTENTION BASED METHODS. CPA IS PLUGGED INTO A SINGLE RESIDUAL BLOCK (CPA_$c2$, CPA_$c3$, CPA_$c4$) RESPECTIVELY AND ALL THREE BLOCKS (CPA_all) AT THE SAME TIME. * DENOTES THAT THE ATTENTION MODULE IS REPRODUCED BY US ON THE SAME BASELINE. THE MAP AND RANK-1 ACCURACY (%) ARE REPORTED.

Method	MARS		Duke-VideoReID	
	mAP	R1	mAP	R1
Baseline	77.7	83.8	89.0	91.3
STA* [17]	78.8	84.7	90.4	92.6
TA* [21]	78.0	84.0	89.7	92.2
TSA* [74]	77.7	85.1	89.4	91.7
MSTA* [75]	78.8	84.9	90.0	92.6
CPA_c2	78.3	84.8	90.0	92.3
CPA_c3	78.8	84.9	90.4	92.3
CPA_c4	78.8	84.5	90.7	92.6
CPA_all	**80.1**	**86.3**	**91.8**	**93.2**

D. Ablation Study

To validate the effectiveness of our proposed method, we conduct extensive ablation experiments to investigate the improvement brought by each component of our framework,

Effectiveness of Each Component. To evaluate the effectiveness of each component in our approach, we conduct several analytic experiments on MARS and Duke-VideoReID datasets. The results are shown in Table IV. We set our baseline model **B** to be Resnet50 without multi-level CPA as backbone network and single head with two baseline loss in section III-B1 as supervision. **B w/o TL** indicates training the baseline model without hard triplet loss. Compare to **B**, our multi-level CPA module in **B+CPA** improves mAP and

TABLE IV
COMPONENT ANALYSIS OF THE PROPOSED APPROACH ON MARS AND DUKE-VIDEOREID DATASETS. B REPRESENTS THE BASELINE MODEL TRAINED ONLY WITH IDENTITY CLASSIFICATION LOSS AND HARD TRIPLET LOSS (TL). CPA IS OUR PROPOSED MULTI-LEVEL CONTEXT-AWARE PART ATTENTION MODULE, FL IS THE MULTI-FRAME CONSISTENCY REGULARIZATION LOSS, MHC IS THE MULTI-HEAD COLLABORATIVE LEARNING FRAMEWORK. HL PRESENTS THE MUTLI-HEAD CONSISTENCY REGULARIZATION LOSS FOR EACH HEAD. DS REPRESENTS THE DENSE SAMPLING STRATEGY DURING EVALUATION STAGE. THE MAP (%) AND RANK-1, -5, -10, -20 ACCURACY SCORES (%) ARE REPORTED.

Method	MARS					Duke-VideoReID				
	mAP	R1	R5	R10	R20	mAP	R1	R5	R10	R20
B w/o TL	75.0	82.4	93.0	95.5	96.8	87.5	90.6	97.3	98.6	99.0
B	77.7	83.8	94.7	96.0	97.2	89.0	91.3	97.9	98.4	99.0
B+FL	78.9	84.9	95.3	96.7	97.4	90.6	92.0	98.9	99.1	99.4
B+CPA	80.1	86.3	95.2	96.7	97.9	91.8	93.2	98.3	99.0	99.7
B+CPA+FL	80.7	86.8	95.4	97.1	97.8	92.6	93.7	98.7	99.0	99.6
B+CPA+FL+MHC	82.7	87.5	96.1	97.6	98.2	94.4	95.3	98.9	99.4	99.6
B+CPA+FL+MHC+HL	83.4	88.2	96.4	97.6	98.3	94.7	96.0	98.9	**99.6**	99.7
B+CPA+FL+MHC+HL+DS	**84.1**	**88.2**	**96.6**	**97.8**	**98.5**	**95.8**	**96.6**	**99.1**	99.4	**99.7**

Rank-1 accuracy by 2.4% and 2.5% respectively on MARS, as well as 2.8% and 1.9% on Duke-VideoReID. These prominent improvement demonstrates that the CPA module are very useful to boost re-ID performance by considering the multiple level features in different convolutional layers. **FL** means the multi-frame consistency loss, which further brings about 0.7% improvement in both mAP and Rank-1 accuracy compared to B+CPA by restricting the ambiguity among frame-level features and weakening the influence of noisy frames. **MHC** refers to the multi-head collaborative learning framework. The multi-head learning framework brings huge improvement. We then add the multi-head consistency loss **HL** to each head to demonstrate the benefit from the knowledge transferring and regularization provided by the prediction consensus of head population, which results 0.7% improvement in both mAP and Rank-1 accuracy in MARS dataset. Both the multi-frame consistency loss and multi-head consistency loss contribute to the overall performance gain on two datasets. This demonstrates that the regularization process is really helpful for training a video re-ID model. By further employing the dense sampling (**DS**) strategy during testing, our approach fully utilizes the information from the whole original video sequence and achieves a very high performance. The **DS** strategy usually requires much more time than the **FS** strategy as discussed in Sec. IV-B.

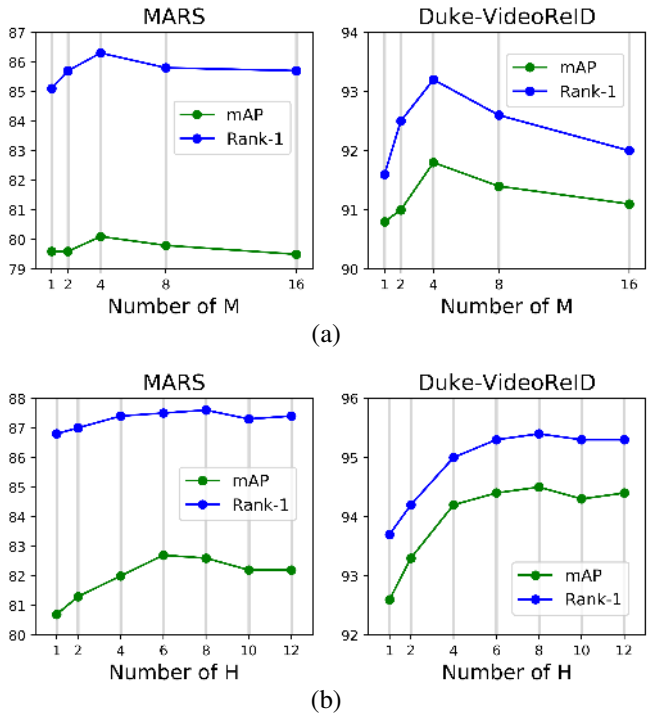

Fig. 6. (a) Performance comparison between variants of **B+CPA** with different number of spatial regions M. (b) Performance comparison between variants of **B+CPA+FL+MHC** with different number of collaborative learning heads (H). The mAP (%) and Rank-1 accuracy (%) on MARS and Duke-VideoReID datasets are reported.

Effectiveness of Multi-Level Attention. We first evaluate the multi-level attention strategy on MARS and Duke-VideoReID datasets in Table III. Compared to the widely used single layer attention module, the proposed multi-level attention method achieves better performance consistently on both datasets. The superiority verifies the proposed idea to utilize the multi-level features in different convolutional layers. Besides, we also demonstrate the superiority of our method in camparison with STA [17], which only employs spatial-temporal attention before final global average pooling layer. Both STA [17] and our proposed method utilize the spatial-temporal attention. However, rather than simply using it as a weighted pooling strategy in *pooling layer* [17], we introduce a multi-level attention module by exploring the attention cues in *multiple different convolution layers* to capture discriminative local features. Results in Table III demonstrate that our proposed strategy consistently performs better than STA [17], typically with 2% improvement in both mAP and Rank-1 accuracy. Several other attention based video re-ID methods are also reproduced and compared, including Temporal Attention (TA) [21] and Temporal Self Attention (TSA) [74] , Multi-scale Spatial-Temporal Attention (MSTA) [75]. All of these attention mechanisms, including CPA, are implemented on the same baseline for fair comparison. These attention mechanisms also only pay attention to the discriminative frame or patch at the end of CNN for feature aggregating. Our multi-level CPA model performs significantly better than other attention mechanisms. The experimental results demonstrate our multi-level CPA has stronger capability to model spatial-temporal information and learn discriminative features.

TABLE V
ANALYSIS OF DIFFERENT HYPER-PARAMETERS ON MARS AND DUKE-VIDEOREID DATASETS. THE MAP AND RANK-1 ACCURACY (%) ARE REPORTED.

Parameter		MARS		Duke-VideoReID	
		mAP	R1	mAP	R1
r	1	**83.4**	88.1	94.5	95.9
	2	**83.4**	**88.2**	**94.7**	**96.0**
	4	82.6	87.8	94.2	95.6
β	0.1	82.9	87.6	94.2	95.7
	0.3	**83.4**	**88.2**	**94.7**	**96.0**
	0.5	83.1	87.9	93.9	95.2
γ	1	83.1	88.0	94.4	95.8
	5	**83.4**	**88.2**	**94.7**	**96.0**
	10	82.5	87.4	93.7	95.1

E. Parameter Analysis.

In this section, we conduct experiments on MARS and Duke-VideoReID dataset to analyze the effect of different parameter settings of CPA and MHC modules respectively: the number of spatial region M in CPA, the number of head H in MHC, and the number of hyper-parameter.

Different number of spatial region M. We analyze the effect with different spatial regions M in CPA module on two datasets. As shown in Fig. 6(a), five different numbers of M: 1, 2, 4, 8 and 16 are evaluated. Here, $M = 4$ corresponding to the **B+CPA** model in Table IV. Notice that when $M = 1$, the CPA module only models the temporal dependency of frame-level features rather than part-level ones. As M increases, the CPA module is able to capture more fine-grained context-aware part attention and aggregate the spatial-temporal context information more flexibly. However, a large M means each region can be too small to contain enough information to present body part, and the relationship between local parts is too complicated to learn. Empirically, we choose $M = 4$ in all the experiments.

Different number of head H. We evaluate the performance with different head numbers H (from 1 to 12) on two datasets, as shown in Fig. 6 (b). Notice that $H = 6$ corresponding to the **B+CPA+FL+MHC** model in Table IV. It is shown that the performance of our model first increases when the head number H increases, and it achieves the best when $H = 6$ or $H = 8$. However, when the head number H continues increasing, the performance begins falling slightly, which may be caused by the increasing variance of the gradients to the backbone network.

Different number of hyper-parameter We conduct a set of experiments on three different hyper-parameters r, γ, and β on two datasets, as reported in Table V. When the channel dimension of CPA keeps constant, *i.e.* $r = 1$, the performance is similar to our setting $r = 2$ that reduces the running memory. However, the smaller dimension makes the performance drop like $r = 4$. Besides, the table shows that the performance is insensitive to the weights β and γ of loss function except for $\gamma = 10$. This phenomenon means that the higher multi-frame consistency loss brings unstable learning and it should be carefully set. We choose the setting $\beta = 0.3$ and $\gamma = 5$ according to the best results.

TABLE VI
ANALYSIS OF DIFFERENT NORMALIZATION METHODS AND MULTI-HEAD CONSISTENCY LOSS FUNCTIONS ON MARS AND DUKE-VIDEOREID DATASETS. THE MAP AND RANK-1 ACCURACY (%) ARE REPORTED.

Method	MARS		Duke-VideoReID	
	mAP	R1	mAP	R1
ℓ_2 Norm	83.2	87.9	94.1	95.5
Softmax Norm	**83.4**	**88.2**	**94.7**	**96.0**
Mean Square Error	82.7	87.5	93.4	94.8
Cross Entropy	**83.4**	**88.2**	**94.7**	**96.0**

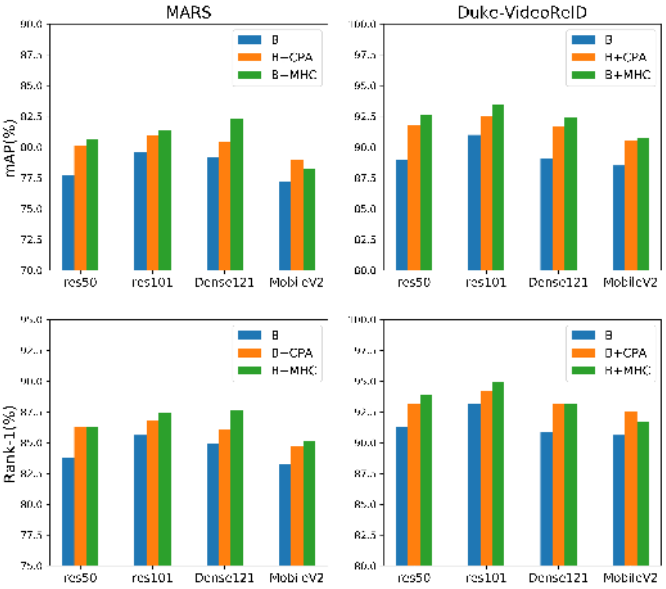

Fig. 7. Performance analysis of CPA and MHC with different backbones (ResNet50, ResNet101, DenseNet121 and MobileNetV2) on MARS and Duke-VideoReID datasets. The mAP (%) and Rank-1 (%) are reported.

F. Normalization and Multi-head Consistency Loss

We also investigate the performance of different normalization methods and multi-head consistency loss functions on two datasets, as shown in Table VI. For the feature normalization, the softmax normalization performs better than the ℓ_2 normalization method, meaning better discriminability. As the identity loss uses the cross-entropy loss to learn features, we naturally use it in multi-head consistency loss. For experiment completeness, we also check the mean square loss [76], showing that the cross-entropy loss is more suitable for multi-head consistency due to its performance improvement.

G. Backbone Analysis.

To demonstrate the generality of our re-ID method, we conduct more experiments to validate the effectiveness of both multi-level CPA model and multi-head collaborative learning scheme in several different backbone architectures, including ResNet50, ResNet101, DenseNet121 and MobileNetV2. The comparison results are shown in Fig. 7. Botch multi-level CPA model and multi-head collaborative learning scheme brings consistent improvement in these different architectures.

H. Application of Other Collaborative Learning Methods.

We also implement another two state-of-the-art collaborative learning methods, EnsembleNet [77] and Feature Fusion

TABLE VII
COMPARISON WITH THE STATE-OF-THE-ART COLLABORATIVE LEARNING METHODS APPLIED ON VIDEO RE-ID TASK. RANK-1 ACCURACY AND MAP (%) ARE REPORTED ON MARS AND DUKE-VIDEOREID DATASETS.

Datasets	MARS		Duke-VideoReID	
Methods	mAP	R1	mAP	R1
EnsembleNet [77]	80.9	86.9	92.0	93.2
FFL [78]	80.9	87.0	91.2	92.7
MHC+FL+HL	81.4	87.1	93.6	94.4

Learning (FFL) [78], and apply them on the same video re-ID baseline model for fair comparison with our multi-head collaborative learning framework. The comparison results are shown in Table VII. EnsembleNet is also constructed from a multi-head supervision structure, but it adopts the co-distillation loss function to jointly train multiple classifier heads. FFL employs two sub-networks to extract the feature representation respectively and then adopt a fused classifier to supervise the backbone network collaboratively with other individual classifiers. Both EnsembleNet and FFL bring significant improvements on two datasets compared to the baseline model. The performance of EnsembleNet and FFL can be further improved after more effort on tuning the coefficient of the distillation loss and structure. The experimental results indicate that the collaborative learning based methods are really helpful to improve the performance.

I. Part Attention Visualization.

We show the learned spatial and temporal attention weights for different parts on MARS dataset in Fig. 8. It is obtained by first averaging the feature maps along the channel dimension, then applying softmax to normalize the sum of value in each body part. The attention weights reflect the quality of each part region, and our approach is robust to noisy frames such as occlusion, pose variation, background clutter and spatial misalignment. Specifically, our model pays less attention to the self-occlusion frame caused by pose variation in the first row. Moreover, in the second row, thanks to the learned context-aware knowledge in early stage of convolutional neural network, our model eliminate effect of the noisy frame rather than focus on the girl with black backpack, where the target man in white shirt is partially occluded by another girl.

J. Efficiency Analysis.

We adopt the floating-point operations (FLOPs) in number of multiply-adds and the required GPU memory during training to measure the computational cost of CNN model. Both the FLOPS and GPU memory of our method is listed in Table VIII. The input to the network is a single tracklet of 4 frames with spatial resolution 256×128. The classification layers are ignored when calculating FLOPS, since they are not used in the inference stage. The batch size is set to be 32 when measuring the GPU memory during training. Both our proposed CPA and MHC increase less than 1% computational cost for FLOPS compared to the baseline model. There is only 1.4% increment of GPU memory when training our model.

We also compare the efficiency and effectiveness between Non-Local block [55] and our CPA module. We simply replace all the CPA modules with Non-Local blocks to realize **baseline+NL** for fair comparison. The CPA module achieves better performance with much less FLOPS and GPU memory compared to Non-Local block. It indicates that our CPA module is a more efficient and effective attention mechanism for video re-ID task compared to Non-Local structure. The comparison results demonstrate that the performance is significantly improved by our approach, due to the excellent network structure. Considering that our CPA module is both time-efficiency and memory-efficiency, it will bring more improvement by inserting more CPA modules in the backbone network as our future work.

TABLE VIII
EFFICIENCY ANALYSIS OF MULTI-LEVEL CPA AND MHC FRAMEWORK. NL INDICATES THE NON-LOCAL BLOCKS. FLOPS DENOTES THE FLOATING POINT OPERATIONS. MEM. DENOTES THE REQUIRED GPU MEMORY DURING TRAINING. INC. DENOTES THE RELATIVELY INCREMENT. MAP (%) ON MARS DATASET IS REPORTED.

Method	FLOPS	Inc.	Mem.	Inc.	mAP	Inc.
Baseline	10.81G	-	9.49G	-	77.7	-
Baseline+CPA	10.82G	0.1%	9.54G	0.5%	80.1	3.1%
Baseline+MHC	10.90G	0.8%	9.58G	0.9%	80.6	3.7%
Baseline+NL [55]	14.05G	29.9%	23.92G	152.1%	78.9	2.0%
Ours	10.91G	0.9%	9.62G	1.4%	83.4	7.3%

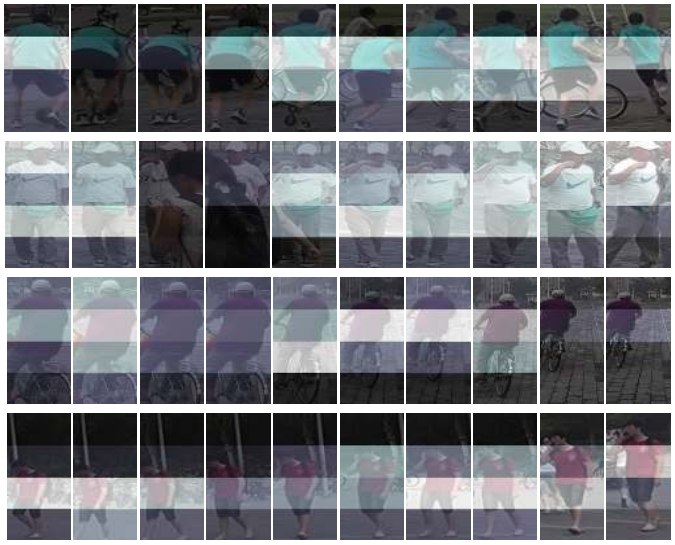

Fig. 8. Visualization of the learned spatial and temporal part attention weights for different parts. The brighter the part, the higher the weight. The four pedestrian tracklets are sampled from MARS dataset.

K. Generalization Ability Analysis.

Different re-ID datasets are usually collected under different environments and introduce bias of data distribution. The model trained on one dataset has a drop in performance when evaluated on other datesets. The cross-dataset performance can indicate the performance when applying a method in real surveillance system. In addition, it can also evaluate the generalization ability of model. To further investigate the generalization ability of our full approach, we conduct cross-dataset experiment between PRID2011 and iLIDS-VID datasets. We compare our full approach with other state-of-the-art works, including CNN-RNN [13], ASTPN [16], TRL [79]

TABLE IX
EVALUATION OF OUR FULL APPROACH WITH OTHER STATE-OF-THE-ART METHODS UNDER CROSS-DATASET SETTING. THE RANK-1, -5, -10, -20 ACCURACY SCORES (%) ARE REPORTED.

Method	PRID2011 to iLIDS-VID				iLIDS-VID to PRID2011			
	R1	R5	R10	R20	R1	R5	R10	R20
CNN-RNN [13]	-	-	-	-	28.0	57.0	69.0	81.0
ASTPN [16]	-	-	-	-	30.0	58.0	71.0	85.0
TRL [79]	8.9	22.8	-	48.8	29.5	59.4	-	82.2
SCAN [61]	9.7	27.5	36.9	48.6	42.8	71.6	80.2	88.9
Ours	**19.3**	**33.3**	**42.7**	**57.3**	**51.7**	**76.4**	**86.5**	**93.3**

and SCAN [61], of which the comparison results is listed in Table IX. The performance drops a lot when evaluting the model on other dataset. When training on PRID2011 dataset, our method achieves 19.3% Rank-1 accuracy. When training on iLIDS-VID dataset, the Rank-1 accuracy of our method is 51.7%. In both cases, our method outperforms other state-of-the-art methods by a large margin. The results suggest that our proposed approach has more powerful generalization ability in comparison with other methods.

V. CONCLUSION

In this work, we have proposed a novel multi-level context-aware part attention model to tackle the video-based person re-ID problem by exploiting the informative features in different convolutional layers. The context-aware part attention module extracts robust and discriminative part features by considering the context information in both spatial and temporal domains. To further improve the performance, we propose a new multi-head collaborative learning scheme with two novel consistency regularization terms. The multi-head collaborative learning framework improves the generalization ability of backbone feature network by stronger supervision from a resemble of learning heads initialized differently. We conduct extensive experiments to demonstrate the effectiveness of each component in our method. The experimental results demonstrate that our video-based person re-ID approach achieves superior performance comparing to previous state-of-the-arts. We are exploring to design a light-weight architecture to capture more precise and informative regions for better person re-ID in the future work.

REFERENCES

[1] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2360–2367.

[2] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai, "Person re-identification by camera correlation aware feature augmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 2, pp. 392–408, 2017.

[3] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1239–1248.

[4] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2288–2295.

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

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

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

[8] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 407–419, 2019.

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

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

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

[12] Y. Yan, B. Ni, Z. Song, M. Chao, Y. Yan, and X. Yang, "Person re-identification via recurrent feature aggregation," in *European Conference on Computer Vision*. Springer, 2016, pp. 701–716.

[13] N. Mclaughlin, J. Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1325–1334.

[14] L. Yu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5790–5799.

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

[16] S. Xu, C. Yu, G. Kang, Y. Yang, and Z. Pan, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4733–4742.

[17] Y. Fu, X. Wang, Y. Wei, and T. Huang, "Sta: Spatial-temporal attention for large-scale video-based person re-identification," in *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2019.

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

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

[20] J. You, A. Wu, L. Xiang, and W. Zheng, "Top-push video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1345–1353.

[21] J. Gao and R. Nevatia, "Revisiting temporal modeling for video-based person reid," *arXiv preprint arXiv:1805.02104*, 2018.

[22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[23] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3741–3750.

[24] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3415–3424.

[25] G. Lisanti, N. Martinel, A. Del, and G. Luca, "Group re-identification via unsupervised transfer of sparse features encoding," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2449–2458.

[26] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2530–2539.

[27] K. Igor, A. Amit, and R. Ehud, "Color invariants for person reidentification," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1622–1634, 2012.

[28] S. Bak and P. Carr, "One-shot metric learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2990–2999.

[29] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3318–3325.

[30] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2288–2295.

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

[32] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3219–3228.

[33] B. Ma, S. Yu, and F. Jurie, "Local descriptors encoded by fisher vectors for person re-identification," in *European Conference on Computer Vision*. Springer, 2012, pp. 413–422.

[34] W. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, 2011, pp. 649–656.

[35] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification." in *Bmvc*, vol. 1, no. 2. Citeseer, 2011, p. 6.

[36] A. J. Ma, P. C. Yuen, and J. Li, "Domain transfer support vector ranking for person re-identification without target camera label information," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3567–3574.

[37] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *European conference on computer vision*. Springer, 2014, pp. 330–345.

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

[39] A. Subramaniam, M. Chatterjee, and A. Mittal, "Deep neural networks with inexact matching for person re-identification," in *Advances in Neural Information Processing Systems*, 2016, pp. 2667–2675.

[40] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[41] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision*. Springer, 2016, pp. 475–491.

[42] W. Li, X. Zhu, and S. Gong, "Person re-identification by deep joint learning of multi-loss classification," *arXiv preprint arXiv:1705.04724*, 2017.

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

[44] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[45] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 274–282.

[46] Y. Lin, Y. Wu, C. Yan, M. Xu, and Y. Yang, "Unsupervised person re-identification via cross-camera similarity exploration," *IEEE Transactions on Image Processing*, vol. 29, pp. 5481–5490, 2020.

[47] M. Ye, A. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsupervised video re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5142–5150.

[48] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2429–2438.

[49] J. Zhou, B. Su, and Y. Wu, "Easy identification from better constraints: Multi-shot person re-identification from reference constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5373–5381.

[50] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6781–6789.

[51] X. Zhu, X. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5683–5695, 2018.

[52] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 388–405.

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

[54] 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 Transactions on Image Processing*, vol. 29, pp. 8821–8830, 2020.

[55] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[56] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1531–1544, 2018.

[57] Z. Liang and J. Shen, "Local semantic siamese networks for fast tracking," *IEEE Transactions on Image Processing*, vol. 29, pp. 3351–3364, 2019.

[58] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European conference on computer vision*, 2018, pp. 3–19.

[59] X. Liao, L. He, and Z. Yang, "Video-based person reidentification via 3d convolutional networks and non-local attention," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 620–634.

[60] J. Li, S. Zhang, and T. Huang, "Multi-scale temporal cues learning for video person re-identification," *IEEE Transactions on Image Processing*, vol. 29, pp. 4461–4473, 2020.

[61] R. Zhang, H. Sun, J. Li, Y. Ge, L. Lin, P. Luo, and X. Wang, "Scan: Self-and-collaborative attention network for video person re-identification," *IEEE Transactions on Image Processing*, 2019.

[62] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[63] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Computer Science*, vol. 14, no. 7, pp. 38–39, 2015.

[64] G. Song and W. Chai, "Collaborative learning for deep neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 1832–1841.

[65] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102.

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

[67] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[68] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3810–3818.

[69] G. Chen, J. Lu, J. Feng, and J. Zhou, "Localized multi-kernel discriminative canonical correlation analysis for video-based person re-identification," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 111–115.

[70] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," in *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2018.
[71] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.
[72] J. Si, H. Zhang, C. Li, J. Kuen, X. Kong, A. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5363–5372.
[73] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.
[74] J. Li, J. Wang, Q. Tian, W. Gao, and S. Zhang, "Global-local temporal representations for video person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3958–3967.
[75] W. Zhang, X. He, X. Yu, W. Lu, Z. Zha, and Q. Tian, "A multi-scale spatial-temporal attention model for person re-identification in videos," *IEEE Transactions on Image Processing*, 2019.
[76] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *arXiv preprint arXiv:1703.01780*, 2017.
[77] H. Li, J. Y.-H. Ng, and P. Natsev, "Ensemblenet: End-to-end optimization of multi-headed models," *arXiv preprint arXiv:1905.09979*, 2019.
[78] J. Kim, M. Hyun, I. Chung, and N. Kwak, "Feature fusion for online mutual knowledge distillation," *arXiv preprint arXiv:1904.09058*, 2019.
[79] J. Dai, P. Zhang, D. Wang, H. Lu, and H. Wang, "Video person re-identification by temporal residual learning," *IEEE Transactions on Image Processing*, vol. 28, no. 3, pp. 1366–1377, 2018.

Dongming Wu is currently working toward the Ph.D. degree in the School of Computer Science, Beijing Institute of Technology, Beijing, China. He received the B.S. degree in Computing Science from Beijing Institute of Technology in 2019. His current research interests include deep learning and computer vision.

Mang Ye received the B.S. and M.S. degrees from Wuhan University, Wuhan, China, in 2013 and 2016, respectively. He is currently a Ph.D student at Department of Computer Science, Hong Kong Baptist University. His research interests focus on multimedia retrieval and computer vision.

Gaojie Lin received the B.S. degree from School of Computer Science, Beijing Institute of Technology in 2017. He is currently a Ph.D student at School of Computer Science, Beijing Institute of Technology, Beijing. His research interests focus on computer vision and video re-identification.

Xin Gao is currently a full Professor of computer science with the Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia. He is also a PI with the Computational Bioscience Research Center, KAUST, and an Adjunct Faculty Member with the David R. Cheriton School of Computer Science, University of Waterloo. He received the the Ph.D. degree in computer science from University of Waterloo, Waterloo, ON, Canada, in 2009. His group focuses on building computational models, developing machine learning methods, and designing efficient and effective algorithms, with particular a focus on applications to key open problems in biology. He has coauthored more than 100 research articles in the fields of machine learning and bioinformatics.

Jianbing Shen (M'11-SM'12) is currently acting as the Lead Scientist at the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. He is also an adjunct Professor with the School of Computer Science, Beijing Institute of Technology, Beijing, China. He has published more than 100 top journal and conference papers. His Google Scholar Citations are more than 10000 times with H-index 51, and twenty-three papers are selected as the ESI Highly Cited Papers. He has also obtained many flagship honors including the Fok Ying Tung Education Foundation from Ministry of Education, and the Program for New Century Excellent Talents from Ministry of Education. His current research interests include computer vision and deep learning. He was rewarded as the receptions of Highly Cited Researcher by the Web of Science in 2020. He serves as an Associate Editor for *IEEE Trans. on Image Processing*, *IEEE Trans. on Neural Networks and Learning Systems*, *Pattern Recognition* and other journals.