IAUnet: Global Context-Aware Feature Learning for Person Reidentification

Ruibing Hou, *Graduate Student Member, IEEE*, Bingpeng Ma, *Member, IEEE*, Hong Chang, *Member, IEEE*, Xinqian Gu, *Student Member, IEEE*, Shiguang Shan, *Senior Member, IEEE*, and Xilin Chen, *Fellow, IEEE*

***Abstract*— Person reidentification (reID) by convolutional neural network (CNN)-based networks has achieved favorable performance in recent years. However, most of existing CNN-based methods do not take full advantage of spatial–temporal context modeling. In fact, the global spatial–temporal context can greatly clarify local distractions to enhance the target feature representation. To comprehensively leverage the spatial–temporal context information, in this work, we present a novel block, interaction–aggregation-update (IAU), for high-performance person reID. First, the spatial–temporal IAU (STIAU) module is introduced. STIAU jointly incorporates two types of contextual interactions into a CNN framework for target feature learning. Here, the spatial interactions learn to compute the contextual dependencies between different body parts of a single frame, while the temporal interactions are used to capture the contextual dependencies between the same body parts across all frames. Furthermore, a channel IAU (CIAU) module is designed to model the semantic contextual interactions between channel features to enhance the feature representation, especially for small-scale visual cues and body parts. Therefore, the IAU block enables the feature to incorporate the globally spatial, temporal, and channel context. It is lightweight, end-to-end trainable, and can be easily plugged into existing CNNs to form IAUnet. The experiments show that IAUnet performs favorably against state of the art on both image and video reID tasks and achieves compelling results on a general object categorization task. The source code is available at https://github.com/blue-blue272/ImgReID-IAnet.**

***Index Terms*— Feature enhancing, interaction–aggregation, person reidentification (reID), spatial–temporal context modeling.**

I. Introduction

PERSON reidentification (reID) aims at retrieving particular persons from nonoverlapping camera views. It has gained increasing attention due to its importance in many applications, such as video surveillance analysis and tracking. Despite much progress has been achieved in recent years [1]–[6], it remains difficult due to tremendous challenges, such as occlusion, background clutters, poses variation, and camera viewpoints variations. Previous approaches mostly focus on image setting, where the persons from different cameras are matched by comparing their still images. With the emergence of video benchmarks [7], [8], the researchers have also started to utilize video data for reID.

Recently, reID by deep neural networks has attracted increasing attention. These approaches utilize convolutional neural networks (CNNs), which typically stacks convolutional and pooling operations, to develop discriminative and robust features. With powerful deep networks and large-scale labeled data sets, CNN-based methods achieve favorable performance and efficiency.

Despite the significant progress in image person reID, most existing CNN-based methods do not take full advantage of spatial context modeling. As point out by Zheng *et al.* [10], the final convolutional features of pedestrians usually focus only on the most representative local regions, which may be indistinguishable for two persons with similar-looking local parts. For example, as shown in Fig. 1(a), the upper clothes of the image pair attract the most attention. However, it is difficult to distinguish the two pedestrians. Varior *et al.* [11] demonstrated that the long-range **global spatial context** can greatly help to clarify local confusions. As illustrated in Fig. 1(b), with the help of the spatial context, the features of upper body parts can be adaptively changed to distinguish the two pedestrians. Therefore, it is desirable to automatically capture the global spatial context for image reID.

For video person reID, current CNN-based methods do not make the best of spatial–temporal context modeling. The 2-D convolution operations completely ignore the temporal information of the video. Although the 3-D convolution [12] operations can capture spatial–temporal context, they are limited to local temporal context modeling [13]. With only the spatial information, the feature generated for a video is often corrupted by the misdetected frames [14]. McLaughlin *et al.* [15] pointed out that the long-range **global temporal context** can help to reduce the interference. As shown in Fig. 1(c),

Manuscript received September 20, 2019; revised February 29, 2020 and June 18, 2020; accepted August 8, 2020. This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61732004, Grant 61876171, and Grant 61976203; and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding author: Bingpeng Ma.)*

Ruibing Hou, Hong Chang, Xinqian Gu, and Xilin Chen are with the Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, and also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: ruibing.hou@vipl.ict.ac.cn; changhong@ict.ac.cn; xinqian.gu@vipl.ict.ac.cn; xlchen@ict.ac.cn).

Bingpeng Ma is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: bpma@ucas.ac.cn).

Shiguang Shan is with the Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China, also with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai 200031, China (e-mail: sgshan@ict.ac.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3017939

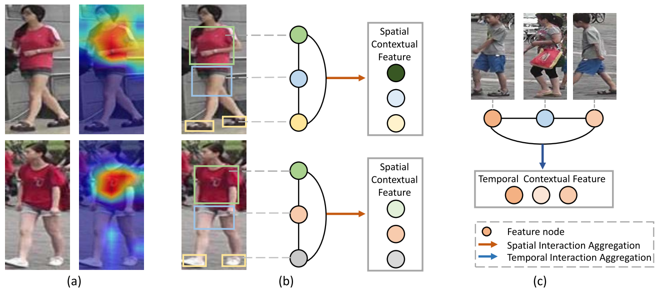

Spatial Contextual Feature
Spatial Contextual Feature
Temporal Contextual Feature
Feature node
Spatial Interaction Aggregation
Temporal Interaction Aggregation
(a)
(b)
(c)

Fig. 1. Illustration of our motivation. (a) Pair of input images and activation maps [9]. The upper clothes attract the most attention. However, they are indistinguishable for the two persons. (b) In spatial context modeling, different body parts interact and aggregate to form structure features with spatial contextual knowledge. Here, the upper body feature can be adaptively updated to distinguish the two persons. (c) In temporal context modeling, the frames interact and aggregate to generate features with temporal contextual information. With the temporal context, the corruption from the misdetected frames can be alleviated.

with the help of temporal context, the features of the misdetected frames can be adaptively updated to describe the target person. Therefore, it is necessary to capture the long-range spatial–temporal context for video reID.

A recent work [16] proposes an interaction-and-aggregation (IA) network for **image reID**. It introduces two modules that aid CNNs in modeling contextual dependencies. One is spatial IA, which models the dependencies between the features of fixed space positions and then aggregates the correlated features belonging to the same body parts. The other is channel IA, where the channel features with similar semantics are aggregated. The incorporation of these modules into a network gives it the ability to adapt its feature representation to contain contextual information.

Following [16], we further propose a unified framework for **image and video reID**. Compared with [16], the major changes of methods are twofold. For one thing, different from [16] that models the spatial dependencies between fixed geometric positions, we go one step further and perform higher level context modeling between disjoint and distant body parts regardless of their shapes. This is conductive to capture longer term spatial contextual dependencies. For another, video reID is considered. We design an additional network module to capture the longer range temporal contexts that can achieve more robust video feature representations.

In this work, we propose a new network module, spatial–temporal interaction–aggregation-update (STIAU), to jointly consider both globally spatial and temporal contexts of the target person. To be specific, given a video feature map, a sequence of intermediate convolutional feature maps of all frames, STIAU generates a spatial–temporal relation map. The relation map captures two types of interactions between body parts: spatial interactions that model the dependencies between disjoint and distant body parts of a single frame and temporal interactions that model the dependencies between the body parts with the same semantics across all frames. In this way, the long-range spatial–temporal context of the video is captured. Based on the relation map, the features of different parts across all video frames are aggregated to generate a spatial–temporal contextual representation. Finally, the spatial–temporal context is incorporated into each frame to form a structured spatial–temporal feature.

Similar to STIAU in principle, the channel interaction–aggregation-update (CIAU) is proposed to further enhance the feature representation via modeling the semantic contextual interactions between the channels of the video feature maps. Especially, for small-scale body parts that easily fade away in the high-level features from CNNs, CIAU can selectively aggregate the semantically similar features across all channels to update and manifest their feature representations.

Both modules are computationally lightweight and impose only a slight increase in model complexity. They can be integrated into an interaction–aggregation-update (IAU) block and readily inserted into CNNs at any depth. In this work, we add IAU block to ResNet-50 [17] to generate IAU Network (IAUnet) for person reID. To demonstrate the universality of the IAU block, we also present results beyond ResNet-50, indicating that the proposed modules are not restricted to specific network architecture.

The contributions of this article are summarized as follows: 1) we propose a unified network, IAUnet, for both image and video person reID; 2) we formulate an STIAU block for learning context-aware features; it designs the **interaction and aggregation** operations that can efficiently capture the **long-range and global context**; and 3) we propose a CIAU block to model the contextual interactions between feature channels. It can further enhance the feature representation by aggregating the semantically similar features. To the best of our knowledge, we are the first to jointly exploit the spatial, temporal, and channel contexts in reID. Experiments on five reID benchmarks show the superiority of the proposed approach. Moreover, IAUnet is effective on general object categorization tasks as demonstrated on CIFAR-100 [18], showing its potential beyond person reID.

II. Related Work

A. Image Person ReID

Image person reID has very rich literature and can be divided into two classes: traditional and deep learning-based approaches. Traditional solutions generally have two stages: extracting handcrafted features and designing robust metrics. Various handcrafted features have been developed. For metric learning, lots of metric learning techniques have been designed to decide whether two images are matched or not. On the other hand, the success of deep learning in image classification has been inspiring a lot of studies in image person reID [5], [6], [19]–[25]. A line of the work uses the Siamese network that takes image pairs or triplets as the inputs. Li *et al.* [26] input a pair of pedestrian images to a CNN, and the model is trained with a verification loss. Hermans *et al.* [27] further employed a triplet loss. Another line adopts identity classification models. Furthermore, Zhang *et al.* [28] trained the model with a joint triplet and classification loss, which achieves state-of-the-art performance.

Spatial Context Modeling: To handle various challenges in person reID, several algorithms have been proposed to impose spatial structure information on target person appearance modeling. The part-based methods that decompose the target person into several parts have been studied actively. For example, the human parsing method [29], [30], pose detection method [31], [32], and body part specific attention modeling [16], [33], [34] have been designed to localize body parts for part-aligned feature extracting and matching. However, the part-based methods ignore the spatial context between different parts; thus, the similar-looking local parts may lead to wrong retrieval results. Recently, Varior *et al.* [11] employed long short-term memory (LSTM) to model the spatial correlations between different local parts. The work demonstrates that the spatial contextual information is beneficial to enhance the discriminative capability of local features. However, LSTM cannot explicitly model the interactions between local body parts. It causes optimization difficulties that need to be carefully addressed [35]. In contrast, the proposed IAU block is lightweight and simple, specialized to model global context in a computationally efficient manner, and designed to enhance the discriminative power of features.

B. Video Person reID

Video person reID is an extension of image approaches, where the sequential data are used instead of an individual image. Early video reID methods mainly focus on handcrafting video features. The powerful feature learning ability of CNN also inspires its application in video reID. To better distill discriminative information from video, the attention-based approaches are gaining popularity. Liu *et al.* [36] predicted a quality score for each frame to weaken the influence of noisy samples. Zhou *et al.* [37] proposed an RNN-based attention mechanism to select the most discriminative frames from the video. Furthermore, the works [14], [38] employ a spatial and temporal attention layer. However, the attention-based methods usually discard the disturbed frames directly, resulting in the loss of spatial and temporal information of video data. In contrast, we explicitly utilize the spatial–temporal context to alleviate the influence of the disturbed frames, without losing any spatial–temporal information.

Spatial–Temporal Context Modeling: Recently, many methods [22], [39]–[41] propose to exploit temporal context on target sequence appearance modeling. A branch of works [15], [42] uses the optical flow that provides the motion features. However, the optical flow only captures the local temporal context between adjacent frames. Another branch adopts the RNN [15], [37], [38] to explore the long-range temporal context. Nevertheless, they can only capture the temporal contextual relations in the end. Thus, they cannot build a hierarchical structure. Besides, all the abovementioned methods ignore the spatial context of videos. On the contrary, the proposed IAU block captures both spatial and temporal contextual information and can be added to the earlier part of CNNs to build a richer hierarchy.

C. Fine-Grained Visual Categorization

Fine-grained visual categorization aims to discriminate similar subcategories that belong to the same superclass. Since the distinctions among similar subcategories are quite subtle and local, existing methods [43]–[46] usually adopt the features of local parts to represent the images. For example, He *et al.* [43] utilized object and part detectors to extract part features, which is free of using the object and part annotations. Peng *et al.* [44] proposed to use a weakly supervised method to generate the part proposals. Furthermore, the works [45], [46] propose a weakly supervised part selection method with spatial constraints. In this work, we also use a weakly supervised part division unit to extract the body part features for input images. However, different from the abovementioned works, the proposed method further models the relations between different parts. This can greatly help to clarify the local confusion caused by seemingly alike parts of different pedestrians.

D. Spatial–Temporal Context Modeling

Generalization of neural networks to automatically model spatial–temporal contextual relations has drawn great attention recently. Xingjian *et al.* [47] proposed a convolutional LSTM for spatial–temporal sequence forecasting. Ji *et al.* [12] developed a 3-D convolution to capture the motion information for video action recognition. To model long-range dependencies, Wang *et al.* [13] proposed a nonlocal (NL) network to model the similarity relations between any pairs of positions. Wang and Gupta [48] further captured the location and similarity relations between the detected objects for video recognition. Gao *et al.* [49] learned a fixed temporal relation between frames to update the exemplar image for visual tracking. However, the abovementioned methods usually model the contextual relations between the fixed geometric positions. In the proposed approach, we go one step further and perform higher level spatial–temporal contextual modeling between disjoint and distant **body parts** regardless of their shape. Comprehensive empirical results verify the effectiveness of the proposed method.

III. Proposed Method

In this section, we first introduce STIAU and CIAU modules. Then, the IAU block, which integrates the two modules, is illustrated. Finally, we present the overall IAUnet for image and video person reID.

A. STIAU Module

Spatial–temporal context of the target person sequence is crucial for video person reID. However, most existing methods either lack the ability of modeling long-range spatial contextual relationships or overlook the temporal contextual knowledge, resulting in highly sensitive to the distracting objects. To this end, we design an STIAU module to form a structured representation with the spatial–temporal context of the target person sequence.

As shown in Fig. 2, suppose that a convolutional video feature map $F \in \mathbb{R}^{T\times H\times W\times D}$ is given, where T, H, W, and D denote the frame number, the height, the width, and the channel number of the feature map, respectively.

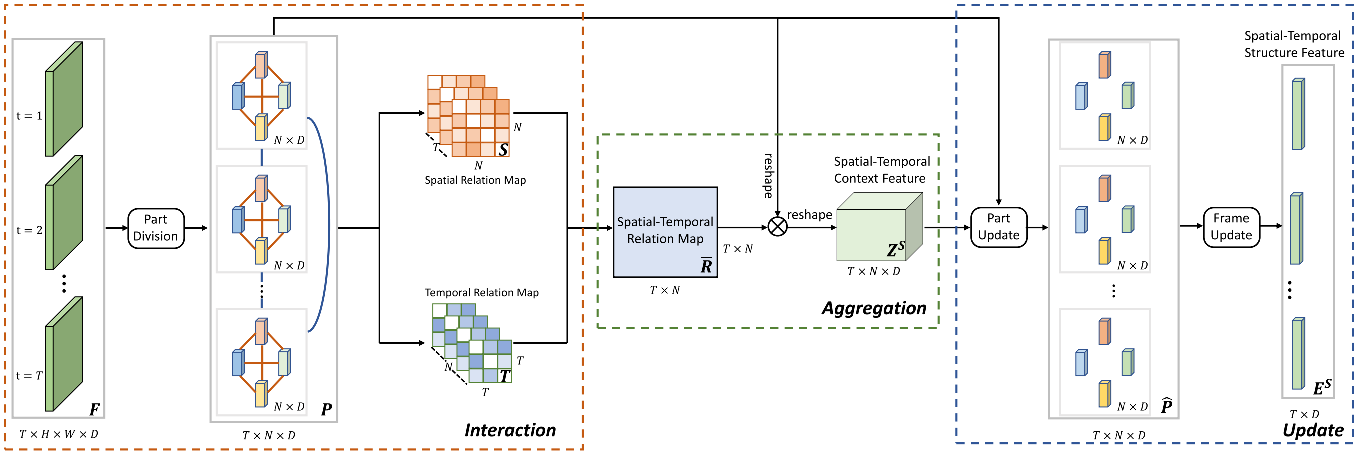


Fig. 2. Architecture of the STIAU module.

We first use a part division unit to extract the part features for each frame. The part features are associated with different body regions, namely, head, upper body, lower body, and shoes. Then, we feed the part features into three sequential operations: interaction, aggregation, and update. Interaction operation explicitly models the dependencies between the parts to produce a spatial–temporal relation map. Two types of relations are considered: spatial relations and temporal relations. The generated relation map is then used to aggregate correlated part features in the following aggregation operation, producing a spatial–temporal context feature. Finally, the context feature is utilized in the update operation to obtain the feature with spatial–temporal structure information.

1) Interaction Operation: As illustrated in Fig. 2, the part division unit takes the video feature map F as inputs and produces the corresponding video part features $P \in \mathbb{R}^{T \times N \times D}$, where N is the number of the parts of each frame. The details of the part division unit will be described later. To perform spatial–temporal context modeling of the target person sequence, the interaction operation models the global contextual relations between the N parts across the T frames, featuring both spatial and temporal relationships.

Especially, we set $P = \{p_{ij} | i = 1, \ldots, T, j = 1, \ldots, N\}$ consisting of all body parts in a person sequence, where $p_{ij} \in \mathbb{R}^D$ is the jth body part feature of the ith frame. Two types of contextual relations are considered: spatial relations and temporal relations. The spatial relations represent the part interactions within the frames. To be specific, we model the spatial relations between any pairs of body parts of each frame, producing a spatial relation map $S = \{S_i\}_{i=1}^T$. Here, $S_i \in \mathbb{R}^{N \times N}$ is the spatial relation map of the ith frame, which is defined as

$$(S_i)_{jk} = W_r^T([|p_{ij} - p_{ik}|, u]), \quad \text{where,} \quad u = \text{GAP}(F) \qquad k \neq j \qquad (1)$$

where $|.|$ denotes the absolute value, $[., .]$ denotes concatenation, and W_r is a weight vector that projects the concatenated vector to a relation scalar. *GAP* stands for global average pooling operation. It performs global spatial–temporal average pooling to the video feature map F to form a coarse global feature of the video, denoted as u ($u \in \mathbb{R}^D$). With the global feature u, the local relations between body parts can be estimated in a global view.

The temporal relations represent the part interactions among frames. In particular, we model the temporal relations of the body parts with the same semantic across all frames. It generates a temporal relation map $T = \{T_i\}_{i=1}^N$. Here, $T_i \in \mathbb{R}^{T \times T}$ is the temporal relation map of the ith body part, which is denoted as

$$(T_i)_{jk} = W_r^T([|p_{ji} - p_{ki}|, u]), \quad \text{where,} \quad k \neq j. \qquad (2)$$

As shown in (2), the coarse global feature u is also used as an input to predict the temporal relations.

Finally, we integrate S and T to form the spatial–temporal relation map $R \in \mathbb{R}^{TN \times TN}$

$$R_{ij} = \begin{cases} (S_{t_1})_{n_1 n_2}, & t_1 = t_2 \\ (T_{n_1})_{t_1 t_2}, & t_1 \neq t_2 \text{ and } n_1 = n_2 \\ 0, & t_1 \neq t_2 \text{ and } n_1 \neq n_2 \end{cases} \qquad (3)$$

where $t_1 = i/N + 1$, $n_1 = i(\text{mod } N) + 1$, $t_2 = j/N + 1$, $n_2 = j(\text{mod } N) + 1$, and R_{ij} denotes the relations between the n_1th part of the t_1th frame and the n_2th part of the t_2th frame. A modified softmax is then used to normalize the relation map

$$\overline{R}_{ij} = \begin{cases} \dfrac{\exp(R_{ij})}{\sum_{k, R_{ik} \neq 0} \exp(R_{ik})}, & R_{ij} \neq 0 \\ 0, & R_{ij} = 0. \end{cases} \qquad (4)$$

Notably, with the decomposition of spatial and temporal relations, each body part is related to $N + T - 1$ parts among a total of NT input parts, which significantly reduces the computation cost of the interaction operation.

2) Aggregation Operation: To make use of $\overline{R}$ in the interaction operation, we follow it with the aggregation operation that aims to aggregate the input video part features based on the relation map. As shown in Fig. 1, we first reshape P to $R^{TN \times D}$ and then perform the matrix multiplication between $\overline{R}$ and P to obtain the spatial temporal context feature $Z^S \in \mathbb{R}^{TN \times D}$

$$Z^S = \overline{R} P. \qquad (5)$$

Z^S is then reshaped to $\mathbb{R}^{T\times N\times D}$ to maintain the size of the input video part feature.

3) Update Operation: With the spatial–temporal context feature, we can compute updated part features using a part update unit. It fuses the initial part feature P and part context feature Z^S to produce the adapted feature $\hat{P}\in\mathbb{R}^{T\times N\times D}$

$$\hat{p}_{ij} = W_{\text{pu}}^{T}\left(\left[p_{ij}, z_{ij}^{S}\right]\right) \tag{6}$$

where $\hat{p}_{ij}\in\mathbb{R}^{D}$ is the jth part feature of the ith frame of $\hat{P}$, and analogously for z_{ij}^{S}, and $W_{pu}\in\mathbb{R}^{2D\times D}$ computes per-part update on the concatenated vector and maintains the input feature dimensionality.

Finally, a frame update unit is applied to each frame to obtain the spatial–temporal structure feature. It first performs global average pooling to $\{\hat{P}_i\}_{i=1}^{T}$ to generate the global feature for each frame. Then, it integrates the frame global feature with the coarse video global feature u to form the spatial–temporal structure feature $E^S\in\mathbb{R}^{T\times D}$

$$E_i^S = W_{\text{fu}}^{T}\left(\left[\frac{\sum_j \hat{p}_{ij}}{N}, u\right]\right) \tag{7}$$

where $W_{\text{fu}}\in\mathbb{R}^{2D\times D}$ computes per-frame update on the concatenated vector and maintains the input feature dimensionality.

4) Part Division Unit: To exploit the local part features for the STIAU module, we should first localize the regions of different body parts. Existing methods [28], [29], [33], [50] usually utilize an external part detection network, making the reID framework too complicated and time consuming. In contrast, we adopt a simple and lightweight spatial attention subnet to localize the body parts. Especially, taking the video feature F as inputs, the subnet uses a convolutional layer to produce the attention maps $A\in\mathbb{R}^{T\times H\times W\times N}$ associated with different body parts

$$A = \sigma(W_a * F + b_a) \tag{8}$$

where σ denotes the sigmoid function, $*$ is the convolutional operation, and $W_a\in\mathbb{R}^{1\times 1\times D\times N}$ and $b_a\in\mathbb{R}^{N}$ are the weights and bias of the convolutional filter. We then generate the video part features P as follows:

$$p_{ij} = \left(\sum_{hw} A_{ihwj} f_{ihw}\right)/(HW). \tag{9}$$

However, the attention maps may focus on background regions. To give a clear clue, we use a body part mask $M\in\mathbb{R}^{T\times H\times W\times N}$ to guide the generation of the attention maps. In detail, first, we use a trained segmentation model [51] to generate the part mask M for input sequences. Then, we resize M to the same size as the attention map. Finally, A and M are flatted to 1-D vectors, respectively. A binary cross-entropy loss is adopted between the flatted A and corresponding flatted M

$$L_p = -\frac{1}{THWN}\sum_{b=1}^{B}\sum_{i=1}^{THWN}\left[M_i(x_b)\log(A_i(x_b)) + (1-M_i(x_b)) \times \log(1-A_i(x_b))\right] \tag{10}$$

where B is the minibatch size, x_b is the bth sequence in the batch, and $M(x_b)$ and $A(x_b)$ are the part mask and attention map of x_b, respectively.

5) Discussion About the Generation of Spatial Relation Map: In the original conference paper [16], the spatial relation map is generated by modeling the semantic similarity between the features of fixed space positions. That is, each position in the feature map is connected with all others and harvests semantically similar contextual information. However, there are two main limits. For one thing, Hou *et al.* [16] used the semantic similarity as the correlation. In general, the features that belong to the same body part have higher semantic similarity than those belonging to different body parts. Thus, Hou *et al.* [16] tend to assign quite low correlations to the positions belonging to different parts, resulting in the lack of the ability to model the dependencies between different body parts. For another, Hou *et al.* [16] need to generate a huge relation map to measure the semantic similarity for all position-pairs of its input. The time and space complexity are both $O(HW\times HW)$, where H and W denote the height and width of the input feature map, respectively. Thus, when the input feature map is with high resolution, SIA [16] would have high computation complexity and take up huge GPU memory.

In the proposed spatial IAU (SIAU)[1] for image reID, the spatial relation map is generated by modeling contextual dependencies between disjoint and distant body parts. Compared with SIA in [16], SIAU has the following advantages.

1) It can perform high-level contextual modeling between different body parts. Different from [16] that mainly models the dependencies within a body part, SIAU uses a subnetwork to predict the correlation between different parts to capture higher level and longer range spatial contextual dependencies. As shown in Fig. 1(b), for the two pedestrians with seemingly similar local body parts, the long-term spatial context can greatly help to clarify the local confusion and, thus, improve the performance.
2) It is with high computational efficiency and GPU memory friendly. Modeling relations between body parts greatly reduces the time and space complexity from $O(HW\times HW)$ to $O(N\times N)$, where N $(N\ll HW)$ is the number of the extracted body parts in each image.

B. CIAU Module

Existing CNN-based methods typically stack multiple convolution layers to extract the features of pedestrians. With the increase of the layer number, the small-scale body parts (e.g., shoes) easily fade away. However, these small-scale parts are very helpful to distinguish the pedestrian pairs with tiny interclass variations. Zhang *et al.* [52] pointed out that most channel maps of high-level features show strong responses for specific parts. Inspired by their views, we build the CIAU module to aggregate semantically similar context features across all channels of a video feature map. By incorporating specific part information from other channel maps, CIAU can enhance the feature representation of that body part.

[1] For image reID, STIAU is equivalent to SIAU since there are no temporal relations for input images.

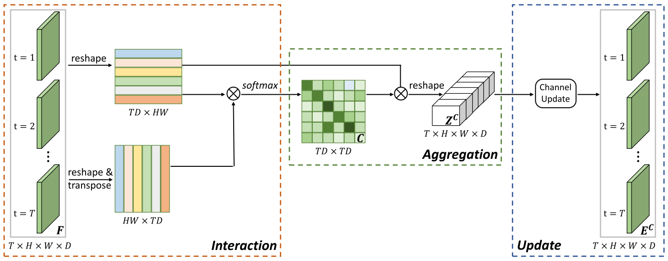


Fig. 3. Architecture of the CIAU module.

1) Interaction Operation: As illustrated in Fig. 3, CIAU module takes a video feature map F as input. In the interaction stage, CIAU explicitly models the semantic contextual relationships between different channels of F to produce a channel relation map. To this end, we first permute and reshape F to $\mathbb{R}^{TD\times HW}$. Then, we perform matrix multiplication between F and the transpose of F and normalize the results to obtain the channel relation map $C \in \mathbb{R}^{TD\times TD}$. Especially, the semantic similarity relation between any two channels is calculated as

$$C_{ij} = \frac{\exp\left(f_i^T f_j\right)}{\sum_{k=1}^{TD} \exp\left(f_i^T f_k\right)} \tag{11}$$

where $f_i, f_j \in \mathbb{R}^{HW}$ denotes the features in the ith and jth channels of F, respectively.

2) Aggregation Operation: Based on the channel relation map, the channel features are then aggregated in the following aggregation operation. To be specific, a matrix multiplication between C and F is performed to obtain the aggregated feature map $Z^C \in \mathbb{R}^{TD\times HW}$

$$Z^C = CF. \tag{12}$$

Z^C is then reshaped and permuted to $R^{T\times H\times W\times D}$ to maintain the input size.

3) Update Operation: We then compute the updated channel features $E^C \in \mathbb{R}^{T\times H\times W\times D}$ based on the aggregated feature map using a channel update unit. It is implemented by a simple convolution layer

$$E^C = W_{\text{cu}} * Z^C + b_{\text{cu}} \tag{13}$$

where $W_{\text{cu}} \in \mathbb{R}^{1\times 1\times D\times D}$ and $b_{\text{cu}} \in \mathbb{R}^D$ are the convolutional filter weights and bias. Note that the resulting feature map aggregates the contexts of different channels according to the channel relation map C. It is complementary to STIAU that aggregates context features of different parts according to the spatial relation map. Similar to STIAU, CIAU can adaptively adjust the input video feature map, helping to boost the feature discriminability.

C. IAU Block Embedding With Networks

We turn STIAU (CIAU) module into STIAU (CIAU) block that can be easily inserted into existing architectures. As shown in Fig. 4, the STIAU (CIAU) block is defined as

$$Y = \text{BN}(E) + F \tag{14}$$

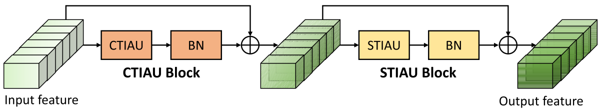


Fig. 4. Architecture of the IAU block.

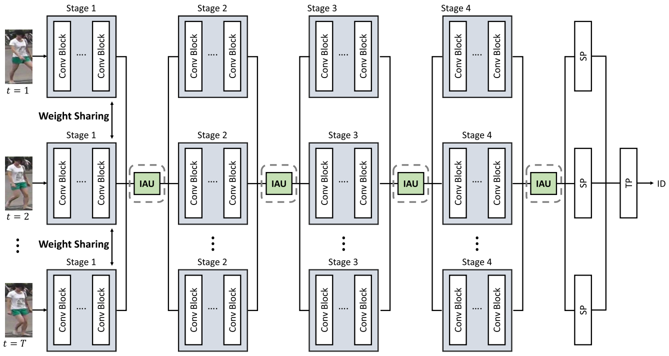


Fig. 5. Architecture of IAUnet for video reID. SP and TP denote spatial pooling and temporal pooling, respectively. When the number of frames in the sequence T is equal to 1, the architecture can be used for image reID.

where F is the input video feature map, E is the output of STIAU or CIAU modules, and BN is a batch normalization layer [53] to adjust the scale of E to the input. A residual learning scheme ($+F$) is adopted along with the IAU mechanism to facilitate the gradient flow. Notably, before entering the BN layer, $E^S \in \mathbb{R}^{T\times D}$ is broadcasted along the spatial dimension to $\mathbb{R}^{T\times H\times W\times D}$ to be compatible with the size of F.

Given an input video sequence, STIAU and CIAU blocks compute complementary contextual relations. We sequentially arrange CIAU and STIAU blocks to form the IAU block, as shown in Fig. 4. IAU block maintains the variable input size and, thus, can be inserted at any depth of networks. Considering the computational complexity, we only place it at the bottlenecks of models where the downsampling of feature maps occurs. Multiple IAU blocks located at bottlenecks of different levels can progressively boost the feature discriminability with a negligible number of parameters.

D. IAUnet for Person ReID

The architecture of IAUnet for person reID is illustrated in Fig. 5. We use ResNet-50 [17] pretrained on ImageNet [54] as the backbone network and modify the output dimension of the classification layer to the number of training identities. Besides, we remove the last spatial downsampling operation in the backbone network, which has been proven to be effective for person reID [34]. We denote the architecture as modified ResNet-50. IAU blocks can be inserted into the backbone network to any stage. Different from previous works [15], [37], [38], [42] that only build temporal contextual dependencies in the end, IAU blocks can capture richer temporal contextual dependencies in the earlier stages. To obtain a single-feature representation for the whole sequence, a temporal average pooling layer is added in the end. Notably, the IAU block can also be used for **image person reID**, where the number of frames in the sequence is set to 1. For image reID, IAU blocks are inserted to stage$_2$ and stage$_3$ of the backbone network. For video reID, we found that a single IAU block added to stage$_2$ gives a comparable result to multiple IAU blocks. Therefore, we only insert an IAU block to stage$_2$ of the backbone network for video reID.

Object Function: Following [55], IAUnet is trained with the combination of classification and ranking. Cross-entropy loss is used for multiclass identifies classification

$$L_{\mathrm{cls}} = -\sum_{b=1}^{B}\left[\log\left(\frac{\exp(p(y=y_b|x_b))}{\sum_k \exp(p(y=k|x_b))}\right)\right] \tag{15}$$

where B is the minibatch size, x_b is the bth input sequence in the batch, y_b is the target label of x_b, and $p(y|x)$ is the probability distribution of predicted label y given input x.

Besides, we adopt a batch hard triple loss [27] for correct ranking. For each sample in a batch, it only selects the hardest positive and hardest negative samples within the batch to form the triples, which is defined as

$$L_{\mathrm{tri}} = \sum_{i=1}^{C}\sum_{a=1}^{K}\left[m + \max_{s=1,\ldots,K} d\big(f_a^i, f_s^i\big) - \min_{\substack{j=1,\ldots,C\\ t=1,\ldots,K\\ j\neq i}} d\big(f_a^i, f_t^j\big)\right]_+ \tag{16}$$

where C is the number of classes (person identities) of the batch, and K is the number of sequences of each class. f_j^i is the final extracted feature corresponding to the jth sequence of the ith person in the batch. d denotes the cosine distance, and m is a margin hyperparameter.

Taking the spatial attention loss in STIAU into consideration, the total loss of IAUnet can be denoted as

$$L_{\mathrm{all}} = L_{\mathrm{cls}} + \lambda_1 L_{\mathrm{tri}} + \lambda_2 L_p \tag{17}$$

where λ_1 and λ_2 are the hyperparameters to balance the effects of different loss functions.

E. Discussion With Other Blocks

In this part, we give a brief discussion on the relations between the proposed IAU block and some existing blocks. The experimental comparisons can be seen in Section IV-D.

1) Relations to Nonlocal: IAU and NL [13] can both be viewed as a graph neural network module. Compared with NL, IAU is more suitable for reID because of the following advantages.

1) NL is a densely connected graph of all spatial positions of input feature maps. It requires computing a dense affinity matrix, which is computationally prohibitive for large-sized feature maps. In contrast, STIAU always has N feature nodes regardless of the size of the input feature map, which is more computationally friendly.
2) NL captures the contextual similarity relations between the fixed geometric positions. STIAU further performs higher level spatial–temporal contextual modeling between disjoint and distant body parts, which can alleviate the local ambiguity to better distinguish similar-looking pedestrians.
3) NL only models the long-range contextual dependencies between spatial features. On the contrary, the proposed CIAU attempts to capture contextual knowledge between the channel features. CIAU is complementary to STIAU and conductive to highlighting important but small details or body parts.

2) Relations to Squeeze-and-Excitation: CIAU has some similarities with squeeze-and-excitation (SE) block [56]. Both blocks are designed to model the contextual dependencies between channels to enhance the feature representation power. However, SE computes the channelwise attention to selectively emphasize informative features, while it is likely to ignore the important but small parts. In contrast, CIAU aggregates the semantically similar contextual features across all channels. It can manifest the feature representations for all body parts.

IV. Experiments

A. Data Sets and Settings

We evaluate the proposed method on three image reID data sets, Market-1501 [57], DukeMTMC [58], and MSMT17 [59], two large-scale video reID data sets, MARS [7] and DukeMTMC-VideoReID [8], and an object category classification data set, CIFAR-100 [18].

Market-1501 is a large-scale data set that contains 1501 identities. The data set is split into two fixed parts: 12 936 images from 751 identities for training and 19 732 images from 751 identities for testing.

DukeMTMC is a subset of the multitarget, multicamera pedestrian tracking data set [60]. There are 36 411 images belonging to 1404 identities. It contains 16 522 training images of 702 identities, 2228 query images of the other 702 identities, and 17 661 gallery images.

MTMC17 is the largest image person reID data set, which contains 126 441 images of 4101 identities. The training set contains 32 621 images of 1041 identities, and the testing set contains 93 820 images of 3060 identities. From the testing set, 11 659 images are randomly selected as query images, and the others are used as gallery images.

MARS is the largest video reID benchmark with 1261 identities and 17 503 sequences captured by 6 cameras. It consists of 631 identities for training and the remaining identities for testing. The bounding boxes are produced by DPM detector and GMMCP tracker, such that it provides a more challenging environment similar to real-world applications.

DukeMTMC-VideoReID is a subset of the tracking data set DuKeMTMC for video reID. The data set consists of 702 identities for training, 702 identities for testing, and 408 identities as distractors. In total, there are 2196 videos for training and 2636 videos for testing.

CIFAR100 is used to show that IAUnet can be also applied to other general recognition problems. This data set contains 60k images of 100 classes with 600 images in each class, where 50k images are used for training and the remaining for testing.

1) Evaluation Metric: We adopt mean Average Precision (mAP) and cumulative matching characteristics (CMC) as evaluation metrics.

2) Part Mask Generating: For each image, we first generate the part masks corresponding to four body parts: head, upper body, lower body, and shoes. To be specific, we first use JPPNet [51][2] pretrained on look into person (LIP) [61] data set to generate the part masks corresponding to 20 semantics.[3] The masks of predictions for different regions are then grouped together to create four coarse labels[4] to guide the part division unit of the STIAU block.

3) Implementation Details for Image ReID: For image reID, the input images are resized to 256 × 128. We use random flipping and random erasing [73] with a probability of 0.5 for data augmentation. The initial learning rate is set to 0.00035 with a decay factor 0.1 at every 20 epochs. Adam [74] optimizer is used with a minibatch size of 64 for 60 epochs training. The margin of triplet loss (m) is set to 0.3, and λ_1 and λ_2 (17) are set to 1 and 0.5, receptively.

4) Implementation Details for Video ReID: For video reID, we randomly sample four frames with a stride of eight frames from the original full-length video to form an input video sequence. The network is trained for 150 epochs in total, with an initial learning rate of 0.0003 and reduced it with a decay rate of 0.1 every 40 epochs. The batch size is set to 32. Other settings and hyperparameters are the same as those in image reID.

5) Implementation Details for Object Classification: For object category classification, we follow the implementation details of MLFN [67]. The initial learning rate is 0.1 with a decay factor 0.1 at every 100 epochs. SGD optimization is used with a 256 minibatch size for 307 epochs training. Other settings are the same as those in image reID. Notably, since there are no specific parts in the CIFIR100 images, the part division unit of STIAU block divides equally the input feature maps into four patches and performs global average pooling on each patch to generate the corresponding part feature.

B. Comparison With the State-of-the-Art Methods

1) Market-1501 and DukeMTMC: In Table I, we compare IAUnet with the state of the art on Market-1501 and DukeMTMC data sets. The compared methods are categorized into three groups, i.e., handcrafted methods, deep learning methods with global features, and deep learning methods with part features. IAUnet achieves the best performance on DukeMTMC and the second-best results on Market-1501. The following is noted.

1) The gaps between our results and those that only employ a global feature [58], [64]–[66], [68], [69], [75] are significant: about 7% mAP improvement. The significant improvements demonstrate that it is effective to employ the spatial contextual information for reID.

TABLE I

COMPARISON WITH THE STATE-OF-THE-ART ON MARKET-1501 AND DUKEMTMC. THE METHODS ARE SEPARATED INTO THREE GROUPS: HANDCRAFTED METHODS (**H**), DEEP LEARNING METHODS ONLY EMPLOYING GLOBAL FEATURES (**G**), AND DEEP LEARNING METHODS EMPLOYING PART FEATURES (**P**), WHERE * DENOTES THOSE REQUIRING AUXILIARY PART DETECTION WHEN TESTING

Methods		Market-1501		DukeMTMC	
		mAP	top-1	mAP	top-1
H	BoW+kissme [57]	20.8	44.4	12.2	25.1
	WARCA [62]	–	45.2	–	–
	LOMO+XQDA [63]	–	–	17.0	30.8
G	SVDNet [64]	62.1	82.3	56.8	76.7
	GAN [58]	66.1	84.0	47.1	67.7
	BraidNet [65]	69.5	83.7	59.5	76.4
	DPFL [66]	72.6	88.6	60.6	79.2
	MLFN [67]	74.3	90.0	62.8	81.0
	KPM [68]	75.3	90.1	63.2	80.3
	Mancs [69]	82.3	93.1	71.8	84.9
	Group [70]	81.6	93.5	69.5	84.9
P	Spindle* [31]	–	76.9	–	–
	PAR [33]	63.4	81.0	–	–
	AACN* [71]	66.9	85.9	59.2	76.8
	PSE* [72]	69.0	87.7	62.0	79.8
	MGCAM [30]	74.3	83.7	–	–
	SPReID* [29]	81.3	92.5	70.9	84.4
	RPP [34]	81.6	93.8	69.2	83.3
	CASN [10]	82.8	94.4	73.7	87.7
	IAnet [16]	83.1	94.4	73.4	87.1
	DSA [28]	87.6	**95.7**	74.3	86.2
	IAUnet	**88.2**	95.0	**79.5**	**89.6**

TABLE II

COMPARISON WITH THE STATE OF THE ART ON MSMT17

Methods	MSMT17			
	mAP	top-1	top-5	top-10
GoogleNet [76]	23.0	47.6	65.0	71.8
Pose-driven [77]	29.7	58.0	73.6	79.4
GLAD [78]	34.0	61.4	76.8	81.6
IANet [16]	46.8	75.5	85.5	88.7
IAUnet	**59.9**	**82.0**	**90.5**	**93.1**

2) Some part-related methods incorporate an external part detection network [28], [29], [31], [72] into the reID model, which makes the reID model too complicated and time-consuming. IAUnet puts much fewer overheads with much better performance on DukeMTMC: about 5% mAP improvement. We argue that the improvement is due to the alleviation of local confusion by modeling the global contextual relations between different body parts.

3) Other attention-centric methods [16], [30], [33], [34], [71] use lightweight attention subnet to learn discriminative body parts. IAUnet outperforms these methods with an improvement of up to 6% on mAP. The superiority of IAUnet over the attention-centric methods further verifies the effectiveness of modeling spatial contextual dependencies among different parts.

2) MSMT17: We further evaluate the proposed method on a recent large scale data set, namely, MSMT17. As shown in Table II, the proposed method significantly outperforms

[2] The code is in https://github.com/Engineering-Course/LIP_JPPNet/

[3] Background, Hat, Hair, Glove, Sunglasses, Upper clothes, Dress, Coat, Socks, Pants, Jumpsuits, Scarf, Skirt, Face, Right-arm, Left-arm, Right-leg, Left-leg, Right-shoe, and Left-shoe.

[4] Head, Upper body, Lower body, and Shoes.

TABLE III

COMPARISON WITH RELATED METHODS ON MARS. THE METHODS ARE SEPARATED INTO TWO GROUPS: DEEP LEARNING METHODS ONLY EMPLOYING GLOBAL VIDEO FEATURES (**G**) AND DEEP LEARNING METHODS EMPLOYING TEMPORAL ATTENTION MODULE (**A**)

Methods		MARS			
		mAP	top-1	top-5	top-10
G	Mars [7]	49.3	68.3	82.6	89.4
	Latent Parts [2]	56.1	71.8	86.6	93.0
	K-reciprocal [79]	68.5	73.9	–	–
	EUG [8]	67.4	80.8	92.1	96.1
	TriNet [27]	67.7	79.8	91.4	–
A	SeeForest [37]	50.7	70.6	90.0	97.6
	Seq-Decision [80]	–	71.2	85.7	91.8
	QAN [36]	51.7	73.7	84.9	91.6
	STAN [14]	65.8	82.3	–	–
	RQEN [81]	71.7	77.8	88.8	94.3
	Snippet [42]	76.1	86.3	94.7	98.2
	TAFD [82]	78.2	87.0	95.4	98.7
	VRSTC [83]	82.3	88.5	96.5	97.4
	IAUnet	**85.0**	**90.2**	**96.6**	**98.3**

TABLE IV

COMPARISONS ON DUKEMTMC-VIDEOREID

Methods	DukeMTMC-VideoReID			
	mAP	top-1	top-5	top-10
EUG [8]	78.3	83.6	94.6	97.6
VRSTC [83]	93.8	95.0	99.1	99.4
IAUnet	**96.1**	**96.9**	**99.5**	**99.8**

TABLE V

OBJECT CLASSIFICATION RESULTS ON CIFAR-100 DATA SET. * INDICATES RESULTS REPORTED BY MLFN [67]

Methods	CIFAR-100
	Error Rates (%)
ResNet-50*	30.21
ResNeXt-50*	29.03
DualNet	27.57
MLFN*	27.21
IAUnet	**20.30**

TABLE VI

COMPARISON TO NL AND SE ON BOTH IMAGE AND VIDEO reID. (a) IMAGE reID DATA SET MARKET-1501. (b) VIDEO reID DATA SET MARS

(a)

Models	Market-1501			
	Param.	GFLOPs	mAP	top-1
baseline	23.5M	4.06	84.5	93.4
NL [13]	26.1M	4.75	86.2	94.2
STIAU	26.1M	4.07	**87.3**	**94.8**
SE [56]	23.7M	4.06	85.4	93.9
CIAU	24.8M	4.86	**86.8**	**94.2**
IAUnet	27.4M	4.87	**88.2**	**95.0**

(b)

Models	MARS			
	Param.	GFLOPs	mAP	top-1
baseline	23.5M	16.24	83.5	88.2
NL [13]	24.0M	19.46	84.0	88.8
STIAU	24.0M	16.25	**84.9**	**89.6**
SE [56]	23.5M	16.24	83.5	88.7
CIAU	23.7M	21.07	**84.5**	**89.1**
IAUnet	24.3M	21.08	**85.0**	**90.2**

existing works [76]–[78] with a top-one accuracy of 20.6% and mAP of 25.9%. IANet models the **intraparts** contextual dependencies to adaptively locate the body parts. IAUnet significantly outperforms it with an improvement of up to 13.1% on mAP, which demonstrates the superiority of **interparts** contextual dependencies modeling.

3) MARS: Table III reports the performance of IAUnet and the current state of the art on MARS. The proposed method outperforms the best existing methods.

1) The works that only employ global video features (**G**) [2], [7], [8], [27], [79] treat each frame of a video equally, resulting in the corruption of the video representation by misdetected frames. IAUnet surpasses these works by up to 10% and 18% on top-one accuracy and mAP, respectively.
2) Other attention-based works (**A**) [14], [36], [37], [42], [80]–[82] leverage a temporal attention network to select the most discriminative frames, resulting in the loss of spatial–temporal information of the video. IAUnet outperforms these works up to 3%. The improvement can be attributed to the feature enhancement by capturing richer spatial–temporal contextual dependencies in IAU blocks.

4) DukeMTMC-VideoReID: As shown in Table IV, the proposed method outperforms the current best result of 1.9% and 2.3% in top-one accuracy and mAP, respectively, on DukeMTMC-VideoReID. VRSTC [83] uses a completion network to recover the appearance of occluded regions as a preprocessing. It is orthogonal to IAUnet and can be easily combined to further improve the performance.

C. Object Categorization Results

In this part, we evaluate IAUnet on a more general object classification task by experimenting on CIFAR-100. Table V compares IAUnet with ResNet-50 [84], ResNeXt-50 [85], DualNet [86], and MLFN [67]. ResNet-50 [84], ResNeXt-50 [85], and MLFN [67] have similar depth and model sizes to IAUnet. The improved result over ResNet-50 shows that IAU blocks bring obvious benefit. IAUnet also outperforms MLFN [67] that fuses multiscale features. Besides, IAUnet beats DualNet [86] that fuses two complementary ResNet branches in an ensemble and has double model size. The consistent improvements suggest that IAU blocks can be easily generalized to general recognition problems.

D. Comparison With Related Approaches

In this section, we present the experimental results compared with NL [13] and SE [56] blocks mentioned in Section III-E. The results are summarized in Table VI. We adopt the modified ResNet-50 model as the baseline and replace the IAU blocks in IAUnet with NL, STIAU, SE, and CIAU blocks representatively.

TABLE VII

PERFORMANCE COMPARISONS OF BASELINE AND PROPOSED SCHEMES ON IMAGE reID TASK

Methods	Market-1501		DukeMTMC	
	mAP	top-1	mAP	top-1
Baseline	84.5	93.4	76.2	87.8
SIA [16]	85.3	93.8	77.1	88.2
STIAU	**86.5**	**94.6**	**78.2**	**89.1**
CIAU	86.3	94.4	78.1	88.9
IAU	**86.7**	**94.8**	**78.9**	**89.2**
IAU (stage$_1$)	85.6	93.8	77.3	88.2
IAU (stage$_2$)	86.7	**94.8**	78.9	89.2
IAU (stage$_3$)	**86.8**	94.3	**79.3**	**89.3**
IAU (stage$_4$)	85.1	93.8	76.7	87.6
IAU-w/o-L_p (stage$_{23}$)	87.3	94.7	78.5	88.9
IAU (stage$_{23}$)	**88.2**	**95.0**	**79.5**	**89.6**

TABLE VIII

PERFORMANCE COMPARISONS OF BASELINE AND PROPOSED SCHEMES ON VIDEO reID TASK

Methods	MARS		Duke-Video	
	mAP	top-1	mAP	top-1
baseline	83.5	88.2	94.5	95.0
STIAU-spatial-only	84.6	88.9	95.3	96.0
STIAU-temporal-only	84.6	89.1	95.0	95.8
STIAU	**84.9**	**89.6**	**95.7**	**96.2**
CIAU	84.5	89.1	95.6	96.0
IAU	**85.0**	**90.2**	**96.1**	**96.9**
IAU-w/o-L_p (stage$_2$)	84.5	88.2	95.3	96.2
IAU (stage$_2$)	**85.0**	**90.2**	**96.1**	**96.9**
IAU (stage$_3$)	84.5	89.1	96.0	96.8
IAU (stage$_{23}$)	**85.3**	90.0	95.9	96.3

TABLE IX

COMPARISONS OF DIFFERENT RELATION CALCULATION STRATEGIES ON MARKET-1501 AND MARS

Methods	Market-1501		MARS	
	mAP	top-1	mAP	top-1
STIAU-L2	84.7	93.5	83.4	88.1
STIAU	**86.5**	**94.6**	**84.9**	**89.6**
CIAU-L2	85.7	94.1	84.0	88.8
CIAU	**86.3**	**94.4**	**84.5**	**89.1**

1) Comparison STIAU With Nonlocal: As shown in Table VI, compared with the NL method, STIAU blocks show higher accuracy under the same model size and less computation budge. We argue that it is hard for NL to directly reason the contextual relations between different body parts. Instead, STIAU blocks are designed to explicitly capture the contextual dependencies between spatially distant parts regardless of their shapes. It can provide complementary features that cannot be easily captured by stacking convolution layers and NL blocks.

2) Comparison CIAU With Squeeze-and-Excitation: As shown in Table VI, we can observe that CIAU achieves better accuracy than SE. We use the default hyperparameters in [56] for SE that leads to marginal improvements. CIAU significantly outperforms SE by 1.4% and 2% mAP on Market-1501 and MARS, respectively. The results indicate that CIAU can better model the contextual interdependencies between channels. The significant improvements also demonstrate that it is more efficient to enhance feature representation power by aggregating similar channel features than multiplying them by constants. Furthermore, by combining CIAU with STIAU, the performance can be further lifted for both image and video reID tasks.

E. Ablation Study

To investigate the effectiveness of each component in the IAU block, we conduct a series of ablation studies on two image reID data sets, Market-1501 and DukeMTMC, and two video reID data sets, MARS and DukeMTMC-VideoReID. Tables VII and VIII summarize the comparison results for different settings. We adopt modified ResNet-50 as the baseline. If there is no special explanation, the proposed blocks are inserted into the last residual block of the stage$_2$ layer of modified ResNet-50.

1) STIAU Blocks: As shown in Tables VII and VIII, STIAU blocks consistently improve the performance remarkably. For image reID, the STIAU block brings about 2% mAP improvements over the baseline. We further compare STIAU to SIA block [16]. As shown in Table VII, STIAU outperforms SIA by about 1% mAP and top-one accuracy, indicating that STIAU can better models the spatial relations. For video reID, we study the effect of STIAU blocks applied along spatial, temporal, and spatial–temporal dimensions. For example, in the spatial-only version, the contextual dependencies only happen within the same frame: i.e., R [in (3)] is simply set to S [in (1)]. Accordingly, the temporal-only version sets R to T [in (2)]. Table VIII shows that both the spatial- and temporal-only versions improve over the baseline, and the performance can be further lifted when the spatial and temporal contextual dependencies are integrated into the STIAU block.

2) CIAU Blocks: We further assess the effectiveness of the CIAU block by adding it to the baseline. CIAU individually outperforms the baseline by about 2% and 1% in terms of mAP and top-one accuracy on image data sets and video data sets, respectively. The improvements indicate that it is effective to enhance feature representation power by aggregating similar features along the channel dimension. When we integrate STIAU and CIAU blocks to the IAU block, the performance can be further improved by about 1% on mAP and top-one accuracy. We argue that the STIAU and CIAU capture the complementary contextual dependencies, spatial, and channel. This leads to each block can provide some complementary features that cannot be easily captured by another block.

3) Relation Calculation Strategy: We exploit another relation calculation strategy, i.e., L2 distance, for STIAU and CIAU blocks. First, Table IX compares STIAU and STIAU-L2, where STIAU and STIAU-L2 use a subnetwork and L2 distance, respectively, to calculate the relations between different parts. We can see that STIAU significantly outperforms STIAU-L2. We argue that the distance metric only models the semantic similarity dependencies, while the subnetwork can model higher level dependencies. For instance, since the head parts typically have highly accurate pedestrian characteristics, the subnetwork can learn to assign high relations between

TABLE X

COMBINING METHODS OF STIAU AND CIAU BLOCKS ON MARKET-1501 AND MARS

Methods	Market-1501		MARS	
	mAP	top-1	mAP	top-1
STIAU+CIAU	85.6	94.3	84.5	88.2
STIAU-CIAU	85.9	94.4	84.4	88.6
CIAU-STIAU	**86.7**	**94.8**	**85.0**	**90.2**

TABLE XI

PERFORMANCE GAIN BY ADDING IAU BLOCKS ON DIFFERENT NETWORKS ON MARKET-1501 AND MARS

Backbone	Method	Market-1501		MARS	
		mAP	top-1	mAP	top-1
ResNet18	baseline	72.0	89.1	74.9	83.7
	ResNet18-IAU	**74.9**	**89.8**	**76.4**	**84.8**
ResNet34	baseline	74.8	89.5	78.4	85.9
	ResNet34-IAU	**78.5**	**91.0**	**79.7**	**86.9**
Inception	baseline	72.2	88.1	72.7	82.2
	Inception-IAU	**75.2**	**89.7**	**74.1**	**83.6**

TABLE XII

PERFORMANCE GAIN BY ADDING IAU BLOCKS ON THE EXISTING reID FRAMEWORK ON MARKET-1501 AND DUKEMTMC

Methods	Market-1501		DukeMTMC	
	mAP	top-1	mAP	top-1
PCB [34]	78.4	91.9	66.1	81.8
PCB-IAU	**79.9**	**92.7**	**70.4**	**83.8**
CAMA [87]	84.5	94.7	72.9	85.8
CAMA-IAU	**87.2**	**95.2**	**78.7**	**89.4**

the head part and other body parts. Then, other body parts can integrate the features of the head part to improve their discrimination. Second, Table IX also compares CIAU and CIAU-L2 where CIAU and CIAU-L2 use dot-product and L2 distance, respectively, to calculate the relations for channel-pairs. The dot-produce and L2 distance both belong to the distance metric, but dot product is more implementation-friendly in deep learning platforms. We can see that CIAU performs slightly better than CIAU-L2.

4) Arrangement of STIAU and CIAU Blocks: We compare three different ways of arranging STIAU and CIAU blocks: sequential STIAU and CIAU blocks (STIAU-CIAU), sequential CIAU and STIAU blocks (CIAU-STIAU), and parallel use of both blocks (STIAU + CIAU). As each block has different functions, the combination mode and order may affect the overall performance. Table X summarizes the experimental results on different arranging methods, where CIAU-STIAU produces the best results on both image and video reID tasks. We argue that STIAU and CIAU blocks compute complementary contextual relations, where STIAU blocks focus on "spatial–temporal" and "different parts" modeling, while CIAU blocks focus on "channel" and "same part" modeling. The CIAU blocks, which can enhance the representation of the individual body part, are conducive to the relationship modeling between different parts by STIAU blocks. Thus, sequential CIAU-STIAU can achieve the best performance.

5) Efficient Positions to Place IAU Blocks: Table VII compares a single IAU block added to the different stages of ResNet50. The block is added to right before the last residual block of a stage. The improvements of an IAU block in stage$_2$ and stage$_3$ are similar but smaller in stage$_1$ and stage$_4$. One possible explanation is that stage$_1$ has a big spatial size that is not very expressive and sufficient to provide precise semantic information. Besides, the visual concepts in stage$_4$ tend to be too abstract; thus, it is difficult to aggregate context features in this stage. Therefore, we only consider adding IAU blocks to stage$_2$ and stage$_3$ layers.

6) Going Deeper With IAU Blocks: Tables VII and VIII also show the results of more IAU blocks. For image reID, IAU blocks can consistently lift the accuracy when more blocks are added. In particular, the model with IAU blocks added to stage$_2$ and stage$_3$ (IAU-stage$_{23}$) improves the model with one IAU block added to stage$_2$ or stage$_3$ (IAU-stage$_2$ or IAU-stage$_3$) by about 1.5 mAP on Market-1501. We argue that multiple IAU blocks can perform hierarchical communication, where each block can provide some complementary relations that cannot be easily captured by other blocks. For video reID, we find that adding two IAU blocks does not give significant gain, as shown in the last three rows of Table VIII. Therefore, we only add a single IAU block to stage$_2$ of the backbone network for video reID.

7) Effect of the Spatial Attention Constrain L_p: To evaluate the contribution of the proposed spatial attention constrain L_p, we train IAUnet and report the results without spatial attention constrain L_p (**IAU-w/o-L_p**). Experimental results are presented in Tables VII and VIII. We can observe that the results of IAUnet consistently outperform that of IAU-w/o-L_p on both image and video reID benchmarks. This confirms the effectiveness of using spatial attention constrain in IAUnet. We argue that without the spatial attention constrain, the learned multiple attention maps tend to be disorganized and focus on the same regions, as shown in Fig. 6(b). Therefore, it is difficult for IAU-w/o-L_p to establish global contextual dependencies between different body parts, resulting in performance degradation.

8) Effectiveness of IAU Blocks Across Different Backbone: We then investigate the generality of IAU blocks on different CNNs. We first investigate the effect of combining IAU blocks with ResNet18 [17] and ResNet34 [17]. The results are summarized in Table XI, where all baseline results are reproduced by ourselves using the same training schema for a fair comparison. We can observe the significant performance improvement induced by IAU blocks. In particular, ResNet18-IAU has an mAP of 89.8% on Market-1501, which is superior to both its direct counterpart ResNet18 (79.1%) as well as the deeper ResNet34 (89.5%). We further exam the effect of IAU blocks on the nonresidual network by experimenting with Inception architecture [76]. We can observe the same phenomena that emerged in the residual architectures. Overall, these experiments demonstrate that IAU blocks can consistently boost the accuracy of a wide range of architectures on both image and video reID tasks.

9) Effectiveness of IAU Blocks Across Existing reID Methods: Finally, we try another two-person reID frameworks,

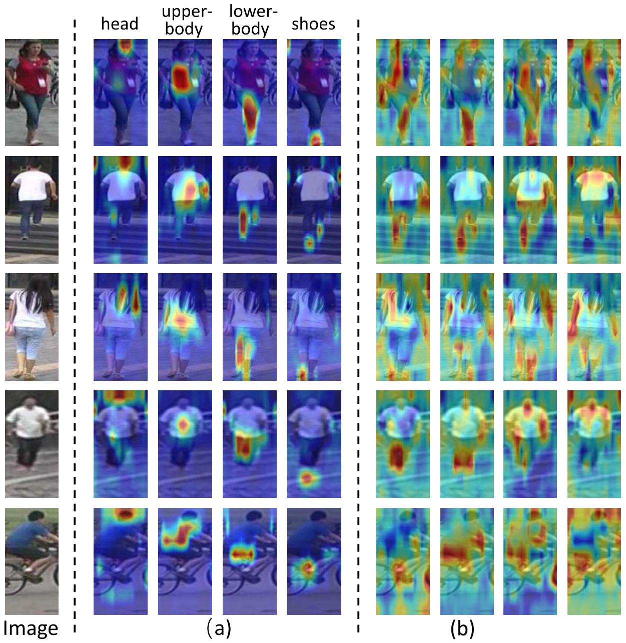


Fig. 6. Learned spatial attention maps. Example images and corresponding receptive fields for part-specific attention maps when $N = 4$. (a) Visualization of the IAUnet. (b) Visualization of the IAUnet trained without spatial attention constrain L_P (IAUnet-wo-L_P).

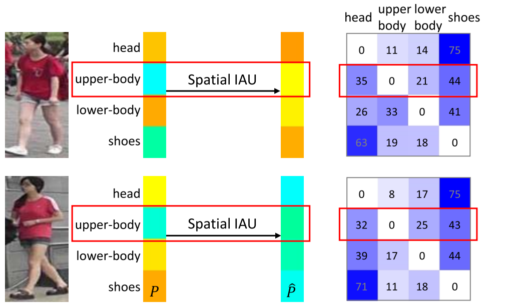
PCB [34] and CAMA [87], to further verify the generality of the proposed IAU block. The results are summarized in Table XII, where the IAU blocks are added to stage$_2$ and stage$_3$ of the backbone of PCB and CAMA to form PCB-IAU and CAMA-IAUnets, respectively. We can observe the significant performance improvement induced by IAU blocks, showing the generality of IAU blocks.

F. Visualizing STIAU Block

In this section, we visualize the learned spatial attention maps of the part division unit in the STIAU block. Fig. 6(a) shows the four spatial attention maps generated by IAUnet for five images. As expected, different spatial attention maps attempt to focus on different local body parts, i.e., head, upper body, lower body, and shoes. It is noteworthy that the spatial attention maps can adaptively localize the body parts under various challenging situations, such as small scale (the second row) and motion blur (the fourth row) even dramatic changes in pose (the last row). Each attention map is used for a weighted average pooling over the whole image, producing a single body-part feature for globally spatial–temporal contextual dependencies modeling. Without the part features, it is really difficult for convolution operations to directly model contextual dependencies between such patterns that might be spatially distant or ill-shaped.

G. Visualizing Relation Maps of STIAU and CIAU Blocks

In this part, we visualize the learned spatial, temporal, and channel relation maps, respectively. Fig. 7 visualizes the initial part features (P in Fig. 2), the updated part features by SIAU ($\hat{P}$ in Fig. 2), and the spatial attention maps (S in Fig. 2) of SIAU. It is clear that, for the two persons with similar upper clothes, the initial upper body features P are difficult to distinguish between the two persons. The spatial relation map S stores global spatial relations. As shown in Fig. 7, for the upper body part, S assigns larger correlation values to the shoes and head parts that are highly discriminative for the input two persons. Therefore, with the feature propagation through S, the upper body features can be updated to distinguish the two persons, as shown in Fig. 7. In addition, we can observe that the head and shoe parts tend to present larger values in S. We argue that the head and shoe parts usually have more accurate pedestrian characteristics than body clothes. Thus, SIAU learns to assign high correlations between the head/shoe parts and other body parts so that the other body parts can integrate the features of the head and shoe parts to improve their discrimination.

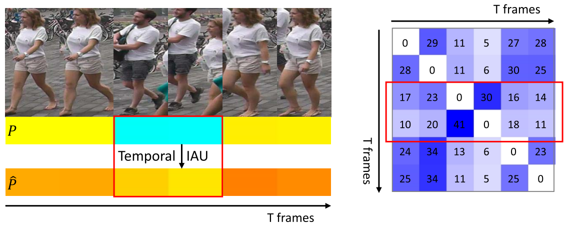


head	upper body	lower body	shoes
0	11	14	75
35	0	21	44
26	33	0	41
63	19	18	0

0	8	17	75
32	0	25	43
39	17	0	44
71	11	18	0

(a) The part features for image pair (b) Spatial relation map S

Fig. 7. Learned spatial relation map. (a) Visualization of the initial part features P and updated part features $\hat{P}$ by SIAU for input image pair. The dimensionality of P and $\hat{P}$ is reduced to $N \times 1$ ($N = 4$) by PCA for visualization. (b) Spatial relation maps S with size $N \times N$.

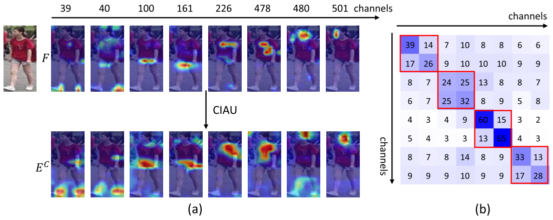


0	29	11	5	27	28
28	0	11	6	30	25
17	23	0	30	16	14
10	20	41	0	18	11
24	34	13	6	0	23
25	34	11	5	25	0

(a) (b)

Fig. 8. Learned temporal relation map. (a) Visualization of the initial part features P and updated part features $\hat{P}$ by temporal IAU for the input sequence. The dimensionality of P and $\hat{P}$ is reduced to $T \times 1$ ($T = 6$) by PCA for visualization. (b) Temporal relation maps T with size $T \times T$.

Fig. 8 visualizes the temporal relation map (T in Fig. 2) for the lower body part of the input sequence. It also visualizes the initial lower body part features (P in Fig. 2) and updated lower body part features ($\hat{P}$ in Fig. 2) by temporal IAU module. It is clear that the detection errors affect the initial part feature, i.e., the feature substantially changes as misdetection happens. The temporal relation map stores the global temporal contextual relations. As shown in Fig. 8, for the misdetected frames, T assigns more than 50% weights to the good frames. Therefore, with the feature propagation through T, the features of misdetected frames can be updated to describe the target person, as shown in Fig. 8. We can also



39	14	7	10	8	8	6	6
17	26	9	10	10	10	9	9
8	7	24	25	13	8	8	7
6	7	25	32	8	9	5	8
4	3	4	9	60	15	3	2
5	4	3	3	13	65	4	3
8	7	8	14	8	9	33	13
9	9	9	10	9	9	17	28

Fig. 9. Learned channel relation map. (a) Visualization of the activation maps of initial channel features F and updated channel features E^C by CIAU for input image. We randomly select eight channels for clearly visualization. (b) Channel relation maps $C \in \mathbb{R}^{8\times 8}$ among the eight channels.

see that the misdetected frames present lower weights in T, indicating their features are suppressed during the aggregation operation and the final video feature is robust to the detection errors.

Fig. 9 visualizes the channel relation map (C in Fig. 3). To visualize the channel relation maps, we randomly select eight channels and visualize their initial features (F in Fig. 3), updated features by CIAU (E^C in Fig. 3), and the relation maps among the eight channels. As shown in Fig. 9, the channel features that focus on the same body parts tend to have a higher correlation. With feature propagation through C, each channel can incorporate the specific part information from other channels. As shown in Fig. 9, the channel feature can focus on more areas of the specific part by CIAU, which enhances its representational power.

V. Conclusion

In this article, we propose an IAU block for globally context modeling that can be effectively implemented by interaction, aggregation, and update operations. The IAU block jointly models spatial–temporal and channel context in a unified framework. We show that by carefully designing the STIAU and CIAU, the proposed IAUnet achieves state-of-the-art results on both image and video reID tasks over some data sets. In the future, we intend to explore a more advanced metric learning approach to further improve performance. Furthermore, we plan to investigate the use of IAU block beyond person reID and object categorization, such as image and video segmentation.

References

[1] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[2] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 384–393.

[3] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[4] X. Liu *et al.*, "HydraPlus-Net: Attentive deep features for pedestrian analysis," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 350–359.

[5] M. Tian *et al.*, "Eliminating background-bias for robust person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5794–5803.

[6] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2530–2539.

[7] L. Zheng *et al.*, "Mars: A video benchmark for large-scale person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 868–884.

[8] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5177–5186.

[9] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[10] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5735–5744.

[11] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang, "A siamese long short-term memory architecture for human re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 135–153.

[12] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[13] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[14] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 369–378.

[15] N. McLaughlin, J. M. del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1325–1334.

[16] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[18] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.

[19] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to rank in person re-identification with metric ensembles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1846–1855.

[20] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2429–2438.

[21] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai, "Hard-aware point-to-set deep metric for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 188–204.

[22] X. Gu, B. Ma, H. Chang, S. Shan, and X. Chen, "Temporal knowledge propagation for image-to-video person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9647–9656.

[23] H. Shi *et al.*, "Embedding deep metric for person re-identification: A study against large variations," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 732–748.

[24] Y. Yang, S. Liao, L. Zhen, and S. Z. Li, "Large scale similarity learning using similar pairs for person verification," in *Proc. AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[25] J. Wu, H. Liu, Y. Yang, Z. Lei, S. Liao, and S. Li, "Unsupervised graph association for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8321–8330.

[26] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.

[27] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[28] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 667–676.

[29] M. M. Kalayeh, E. Basaran, M. Gokmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1062–1071.

[30] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.

[31] H. Zhao *et al.*, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1077–1085.

[32] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose invariant embedding for deep person re-identification," 2017, *arXiv:1701.07732*. [Online]. Available: http://arxiv.org/abs/1701.07732

[33] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3239–3248.

[34] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 480–496.

[35] K. Zolna, D. Arpit, D. Suhubdy, and Y. Bengio, "Fraternal dropout," 2017, *arXiv:1711.00066*. [Online]. Available: http://arxiv.org/abs/1711.00066

[36] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4694–4703.

[37] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6776–6785.

[38] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4743–4752.

[39] J. Li, S. Zhang, and T. Huang, "Multi-scale 3D convolution network for video based person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8618–8625.

[40] R. Hou, H. Chang, B. Ma, S. Shan, and X. Chen, "Temporal complementary learning for video person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–17.

[41] X. Gu, B. Ma, H. Chang, H. Zhang, and X. Chen, "Appearance-preserving 3D convolution for video-based person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 1–16.

[42] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1169–1178.

[43] X. He, Y. Peng, and J. Zhao, "Which and how many regions to gaze: Focus discriminative regions for fine-grained visual categorization," *Int. J. Comput. Vis.*, vol. 127, no. 9, pp. 1235–1255, Sep. 2019.

[44] Y. Peng, X. He, and J. Zhao, "Object-part attention model for fine-grained image classification," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1487–1500, Mar. 2018.

[45] Y. Zhang *et al.*, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1713–1725, Apr. 2016.

[46] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4075–4081.

[47] S. Xingjian, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[48] X. Wang and A. Gupta, "Videos as space-time region graphs," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 399–417.

[49] J. Gao, T. Zhang, and C. Xu, "Graph convolutional tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4649–4659.

[50] X. Qian *et al.*, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 650–667.

[51] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Apr. 2019.

[52] S. Zhang, J. Yang, and B. Schiele, "Occluded pedestrian detection through guided attention in CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6995–7003.

[53] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[54] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.

[55] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[56] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*. [Online]. Available: http://arxiv.org/abs/1709.01507

[57] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[58] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[59] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[60] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 17–35.

[61] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 932–940.

[62] C. Jose and F. Fleuret, "Scalable metric learning via weighted approximate rank component analysis," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 875–890.

[63] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2197–2206.

[64] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3800–3808.

[65] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1470–1478.

[66] Y. Chen, X. Zhu, and S. Gong, "Person re-identification by deep learning multi-scale representations," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 2590–2600.

[67] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2109–2118.

[68] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6886–6895.

[69] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 365–381.

[70] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep CRF for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8649–8658.

[71] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.

[72] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 420–429.

[73] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," 2017, *arXiv:1708.04896*. [Online]. Available: http://arxiv.org/abs/1708.04896

[74] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[75] H.-X. Yu, A. Wu, and W.-S. Zheng, "Cross-view asymmetric metric learning for unsupervised person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 994–1002.

[76] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[77] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," 2017, *arXiv:1709.08325*. [Online]. Available: http://arxiv.org/abs/1709.08325

[78] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global-local-alignment descriptor for pedestrian retrieval," in *Proc. ACM Multimedia Conf. (MM)*, 2017, pp. 420–428.

[79] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

[80] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6781–6789.

[81] G. Song, B. Leng, Y. Liu, C. Hetang, and S. Cai, "Region-based quality estimation network for large-scale person re-identification," 2017, *arXiv:1711.08766*. [Online]. Available: http://arxiv.org/abs/1711.08766

[82] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-S. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4913–4922.

[83] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "VRSTC: Occlusion-free video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7183–7192.

[84] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 630–645.

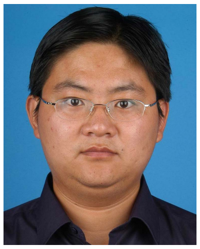
[85] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[86] S. Hou, X. Liu, and Z. Wang, "DualNet: Learn complementary features for image recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 502–510.

[87] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.

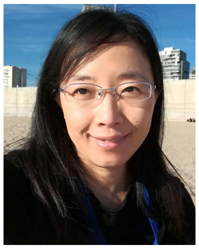
Ruibing Hou (Graduate Student Member, IEEE) received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2016. She is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

Her research interests are in machine learning and computer vision. She especially focuses on person reidentification and few-shot learning.



Bingpeng Ma (Member, IEEE) received the B.S. degree in mechanics and the M.S. degree in mathematics from the Huazhong University of Science and Technology, Wuhan, China, in 1998 and 2003, respectively, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2009.

He was a Post-Doctoral Researcher with the University of Caen, Caen, France, from 2011 to 2012. He joined the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, in March 2013, where he is currently an Associate Professor. His research interests cover computer vision, pattern recognition, and machine learning. He especially focuses on person reidentification, face recognition, and the related research topics.



Hong Chang (Member, IEEE) received the bachelor's degree from the Hebei University of Technology, Tianjin, China, in 1998, the M.S. degree from Tianjin University, Tianjin, in 2001, and the Ph.D. degree from The Hong Kong University of Science and Technology, Hong Kong, in 2006, all in computer science.

She was a Research Scientist with the Xerox Research Centre Europe, Meylan, France. She is currently a Researcher with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her main research interests include algorithms and models in machine learning, and their applications in pattern recognition and computer vision.

Xinqian Gu (Student Member, IEEE) received the B.S. degree in software engineering from Chongqing University, Chongqing, China, in 2017. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China.

His research interests are in computer vision, pattern recognition, and machine learning. He especially focuses on person reidentification, video analytics, and the related research topics.

Shiguang Shan (Senior Member, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004.

He has been a Full Professor with ICT since 2010, where he is currently the Deputy Director of the CAS Key Lab of Intelligent Information Processing. He is also with CAS Center for Excellence in Brain Science and Intelligence Technology, Shanghai, China. He has published more than 300 articles, with totally more than 20000 Google Scholar Citations. His research interests cover computer vision, pattern recognition, and machine learning.

Dr. Shan was a recipient of the China's State Natural Science Award in 2015, and the China's State S&T Progress Award in 2005 for his research work. He has served as the Area Chair of many international conferences, including Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), the Association for the Advance of Artificial Intelligence (AAAI), International Joint Conference on Artificial Intelligence (IJCAI), Asian Conference on Computer Vision (ACCV), International Conference on Pattern Recognition (ICPR), International Conference on Automatic Face and Gesture Recognition (FG), and so on. He has been an Associate Editor of several journals, including the IEEE TRANSACTIONS ON IMAGE PROCESSING (T-IP), *Neurocomputing*, *Computer Vision and Image Understanding* (CVIU), and *Pattern Recognition Letters* (PRL).

Xilin Chen (Fellow, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China. He has authored one book and more than 300 articles in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces.

Dr. Chen is a fellow of the Association for Computing Machinery (ACM), International Association on Pattern Recognition (IAPR), and China Computer Federation (CCF). He has served as an organizing committee member for some conferences, including the General Co-Chair of the IEEE FG13/FG18 and the Program Co-Chair of the ACM International Conference on Multimodal Interaction (ICMI) 2010. He was the Area Chair of Conference on Computer Vision and Pattern Recognition (CVPR) 2017/2019/2020 and IEEE International Conference on Computer Vision (ICCV) 2019. He is also a Senior Editor of the *Journal of Visual Communication and Image Representation* and an Associate Editor-In-Chief of the *Chinese Journal of Computers* and the *Chinese Journal of Pattern Recognition and Artificial Intelligence*.