This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,

GAReID: Grouped and Attentive High-Order Representation Learning for Person Re-Identification

Pingyu Wang, Fei Su, Zhicheng Zhao, Yanyun Zhao, Nikolaos V. Boulgouris

***Abstract*—As person parts are frequently misaligned between detected human boxes, an image representation that can handle this part misalignment is required. In this work, we propose an effective *Grouped Attentive Re-Identification* (GAReID) framework to learn part-aligned and background-robust representations for person re-identification. Specifically, the GAReID framework consists of *Grouped High-Order Pooling* (GHOP) and *Attentive High-Order Pooling* (AHOP) layers, which generate high-order image and foreground features, respectively. In addition, a novel *Grouped Kronecker Product* (GKP) is proposed to employ both channel group and shuffle strategies for high-order feature compression, while promoting the representational capabilities of compressed high-order features. We show that our method derives from an interpretable motivation and elegantly reduces part misalignments without using landmark detection or feature partition. This paper theoretically and experimentally demonstrates the superiority of the GAReID framework, achieving state-of-the-art performance on various person re-identification datasets.**

***Index Terms*—Person Re-Identification, Part Misalignments, Kronecker Product, Group Shuffle, High-Order Pooling**

I. Introduction

PERSON Re-Identification (ReID) aims at matching person images of the same person across non-overlapping cameras. It plays an important role in various video surveillance applications such as suspect tracking and missing elderly or children retrieval. With the blooming of *Convolutional Neural Network* (CNN), the current deep feature learning based methods [1–16] have significantly outperformed a variety of traditional feature learning based approaches [17–22]. However, the ReID task is far from being solved because of part misalignments caused by camera views, detection errors, body occlusions and background clutters. As shown in Fig. 1, part misalignments usually change the spatial distribution of person appearances, which might degenerate the distinctiveness and robustness of person representations.

In order to mitigate part misalignments, prior ReID works have broadly followed two main paradigms, *i.e.*, part-based and landmark-based methods. The part-based approaches [2, 5, 12–14] partition the global person images/features into a few fixed rigid parts and concentrate on local feature learning so as to obviate the need for landmark detection. Nevertheless, such coarse partition is unable to effectively align body parts without considering fine-grained pose variations within each part. For achieving fine-grained part alignments, the landmark-based works [1, 6–11, 23] employ human landmark annotations or landmark detection networks and then learn part-aligned features from pose-normalized person images. Although those works have boosted the ReID performance, they introduce extra operations to the ReID system, *e.g.*, landmark detection and pose normalization. In addition, those operations bring non-ignorable space and time costs, making it hard to train the ReID model.

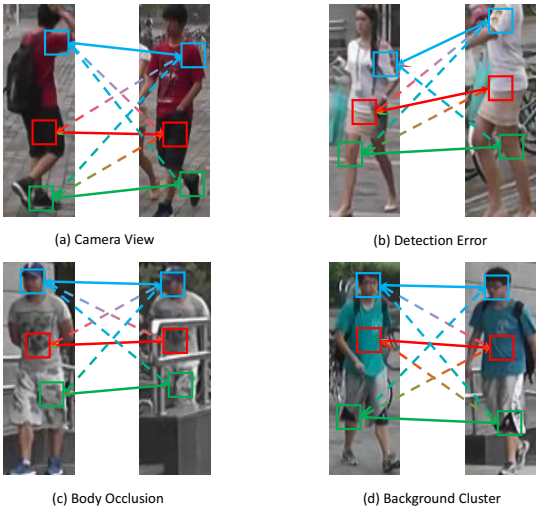


Fig. 1: Illustration of the part misalignment problem caused by camera views, detection errors, body occlusions and background clutters. The aligned part pairs are connected with solid lines, while the misaligned part pairs are connected with dashed lines.

In this work, we propose an effective *Grouped Attentive Re-Identification* (GAReID) framework composed of two novel pooling layers, *i.e.*, *Grouped High-Order Pooling* (GHOP) and *Attentive High-Order Pooling* (AHOP). As we know, compared with the first-order function, the high-order function $f(x) = x^n \, (n > 1, x \geq 0)$ contributes to amplifying the discrepancies between two dependent variables when two independent variables are fixed. Motivated by this amplification property of the high-order function, the essential idea behind GAReID is to compute the high-order mapping of part similarities in order to enlarge the similarity discrepancies between aligned and misaligned part pairs. Specifically, the

Pingyu Wang, Fei Su, Zhicheng Zhao and Yanyun Zhao are with Beijing Key Laboratory of Network System and Network Culture, School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. (e-mail: applewangpingyu@bupt.edu.cn; sufei@bupt.edu.cn; zhaozc@bupt.edu.cn; zyy@bupt.edu.cn)

Nikolaos V. Boulgouris is with Department of Electronic and Computer Engineering, Brunel University, London, United Kingdom. (e-mail: Nikolaos.Boulgouris@brunel.ac.uk)

This work is supported by Chinese National Natural Science Foundation (62076033, U1931202). (Corresponding author: Zhicheng Zhao)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,

GAReID is able to highlight aligned part similarities and suppress misaligned part similarities. Since the high-order feature similarity between a pair of person images is equivalent to an average of high-order similarities of both aligned and misaligned part pairs, the high-order aligned similarities are likely to dominate the high-order feature similarities. In this way, the part misalignment problem is effectively alleviated without relying on landmark detection or feature partition.

Although high-order features contribute to part alignments, the dimension of high-order features increases exponentially, which gravely impairs the applications of high-order models. Therefore, we need to design an effective feature compression method for high-order features. Inspired by light-weight networks [24, 25], the proposed GHOP layer adopts channel group and shuffle strategies to compress the dimension of high-order features. Specifically, input feature channels are uniformly divided into different groups and then those groups are shuffled to disperse the information across feature groups. Subsequently, we propose *Grouped Kronecker Product* (GKP) to employ the Kronecker product for sub-features in each original and shuffled group to excavate informative high-order interactions. Since the Kronecker product increases feature dimensions in each group, we obtain grouped high-order features by conducting element-wise aggregation, which can significantly improve the effectiveness of high-order features. As background clutters may hinder part alignments, we put forward an effective foreground attention module named *Adaptive Foreground Attention* (AFA) to preserve foreground regions and eliminate background areas. With the integration of the GHOP layer and the AFA module, the proposed AHOP layer is constructed to boost both part-aligned and background-robust representation learning.

In summary, this paper makes the following contributions: (1) We analyze the cause of part misalignments and prove that the high-order mapping of part similarities facilitates fine-grained part alignments in theory. (2) We propose an effective GAReID framework with two novel pooling layers, *i.e.*, GHOP and AHOP. The GHOP layer aims at compressing high-order features, while the AHOP layer focuses on eliminating background clutters. (3) The GAReID framework is able to learn both part-aligned and background-robust representations without relying on any landmark detection or feature partition, making it highly generalizable to other unknown pose and background variations. (4) The GAReID achieves state-of-the-art ReID performance on Market1501 [26], CUHK03 [27], DukeMTMC [28] and MSMT17 [29] datasets.

II. Related Works

A. Person Re-Identification

For relieving part misalignments, prior ReID works can be roughly summarized into two streams, *i.e.*, part-based and landmark-based methods. The part-based works [2, 5, 8–10, 12–14, 30, 31] usually use deep neural networks for learning discriminative local features. As global features learned from the full image intend to capture the coarse-grained clues of appearance, the global feature maps in [2, 12, 14] are equally divided into multiple horizontal patches to exploit local details. Based on PCB [2, 12], some following works, *i.e.*, MGN [5], PyramidNet [13] and HPM [14], extract both global and local person representations by dividing convolutional feature maps horizontally into multi-grained patches. To enhance the part alignment of learned representations, the landmark-based works [1, 6–11, 23] consider extra landmark knowledges for training ReID networks. For instance, GAN-based works [6, 7] use auxiliary landmark annotations to guide the generative model [32] to synthesize pose-specific person images and supervise the identity encoder model to mine pose-aligned features. Two-stream networks [33, 34] are applied in [3] to independently generate appearance and pose representations which are fused to enable part alignments. To achieve a more precise alignment, the fine-grained pixel-level person semantics predicted by DensePose [35] are used in [11] as an additional regularizer to guide the part-aligned representation learning from the original images. In order to solve the occluded person ReID problem, Occluded-ReID [36] incorporates the pose information to make the ReID model focus on the body region only and filter noise features brought by occlusions.

In general, these methods use either local feature partition or additional landmark information to align person features. However, the part-based ReID approaches only achieve the coarse-grained part alignments without considering detailed pose variations within each part. In addition, it is non-trivial to obtain landmark-labeled person images or landmark detection networks in real-world circumstances. Therefore, the landmark-based models might not generalize well to new images with unseen pose variations. In this work, the GAReID heads from a totally disparate but effective idea that emphasizes the similarity discrepancies between aligned and misaligned part pairs via a high-order mapping function. Furthermore, our method is able to automatically rectify part misalignments without depending on landmark information or feature partition. Besides, the GAReID framework can be applied to the field of unsupervised person ReID [37], and helps unsupervised person ReID models to select more reliable neighborhoods for each person image. As a result, the proposed GAReID framework has increased practical significance and application prospect.

B. High-Order Statistics

High-order statistics has been widely studied in traditional machine learning due to its powerful representation ability. Recently, the fine-grained visual classification task [38–40] has shown that the integration of high-order features with deep networks can bring promising performance improvements. For person re-identification, Ustinova *et al.* [41] propose an architecture based on the deep bilinear convolutional network. Chen *et al.* [42] construct a mixed-order attention module to utilize both low-order and high-order statistics in attention mechanism, so as to produce discriminative attention proposals. Although the two architectures lead to some performance improvements, they are not explicitly concerned with part alignments.

Although high-order features exhibit strong representational capabilities, the dimension of high-order features exponentially increases, which hinders their applications in real-

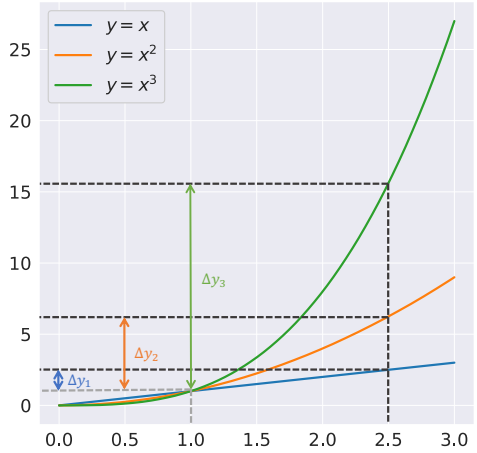


Fig. 2: Illustration of the first-order, second-order and third-order functions. The high-order function contributes to enlarging the difference of dependent variables when the difference of independent variables are fixed, *i.e.*, $\Delta y_3 > \Delta y_2 > \Delta y_1$.

world problems. Recently, several works [39, 40, 43, 44] seek various feature compression methods in order to learn compact high-order features. For example, both CBP [43] and KP [39] adopt random feature projections [45–48], but introduce constant projection matrixes, resulting in additional non-negligible memory overheads. Besides, they employ *Fast Fourier Transform* (FFT) and *Inverse Fast Fourier Transform* (IFFT) to simplify convolution operations, but it may be discommodious to achieve FFT and IFFT on deep learning frameworks. DBT [40] adopts tensor partition to capture intra-group interactions, but ignores inter-group interactions. Moreover, since both CBP and DBT are second-order modules, they are unable to learn higher-order ($n \geq 3$) features, which might severely weaken the generalization capabilities of trained models. In this work, the GAReID employs both channel group and shuffle strategies to achieve high-order feature compression, while promoting the representational capabilities of compressed high-order features.

C. Attention Mechanism

Attention mechanism, inspired by the human sensing process, has been studied extensively in various computer vision tasks [49]. Specifically, an attention mechanism aims at emphasizing informative regions for image representations, while depreciating harmful ones (*e.g.*, background and occluded regions). Interestingly, this approach is also efficient and effective for person re-identification [42, 50–53] because it can handle person misalignments and background clutters. For instance, HACNN [50] jointly learns hard region-level attention and soft pixel-level attention in a unified attention block. Mancs [51] considers both the channel-wise and spatial-wise attention in a fully attentional block, where the channel information is re-calibrated and the spatial structure information is also preserved. In addition, SONA [52] introduces a second-order non-local attention network to directly model long-range relationships via second-order feature statistics. As distinguished from previous attention methods, our AFA module generates foreground attention masks according to the l_2 norms of all spatial features. Interestingly, the AFA module significantly contributes to discovering useful semantic regions without introducing any learnable parameters. By combining the proposed GHOP layer and AFA module, we build the AHOP layer to jointly relieve part misalignments and background clutters.

Theorem 1. *Suppose* $\otimes_n \boldsymbol{u} = \boldsymbol{u} \otimes \boldsymbol{u} \otimes \cdots \otimes \boldsymbol{u}$ *and* $\otimes_n \boldsymbol{v} = \boldsymbol{v} \otimes \boldsymbol{v} \otimes \cdots \otimes \boldsymbol{v}$ *are two* n*th-order vectors generated by Kronecker product* $\otimes$ *with two input vectors* $\boldsymbol{u}$ *and* $\boldsymbol{v}$*, the similarity of* n*th-order vectors is computed by* $\left\langle \otimes_n \boldsymbol{u}, \otimes_n \boldsymbol{v} \right\rangle = \left\langle \boldsymbol{u}, \boldsymbol{v} \right\rangle^n$.

Proof. See the proposition 2 in [54]. □

III. Proposed Method

In this section, we first analyze theoretically the cause of part misalignments. Then we introduce the *Grouped High-Order Pooling* (GHOP) and *Attentive High-Order Pooling* (AHOP) in the GAReID framework as shown in Fig. 3.

A. Part Misalignment

In this part, we give a theoretical analysis for the cause of part misalignments in the person ReID task. Given two input person images $\boldsymbol{I}_u$ and $\boldsymbol{I}_v$ from the same class, we use CNNs in order to extract two convolutional feature maps $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{C \times H \times W}$, where C, H and W denote the channel, height and width dimension, respectively. Then the two feature maps are pooled by a *Global Average Pooling* (GAP) layer [55] to obtain the corresponding person descriptors as follows,

$$\boldsymbol{u} = \frac{1}{|\mathcal{S}|} \sum_{p_u \in \mathcal{S}} \boldsymbol{U}_{p_u}, \ \boldsymbol{v} = \frac{1}{|\mathcal{S}|} \sum_{p_v \in \mathcal{S}} \boldsymbol{V}_{p_v}, \tag{1}$$

where $\boldsymbol{U}_{p_u}, \boldsymbol{V}_{p_v} \in \mathbb{R}^C$ are two part descriptors at the positions p_u and p_v, respectively. The set $\mathcal{S} = \{1, 2, \cdots, HW\}$ is the set of all spatial positions and $|\mathcal{S}| = HW$ is its cardinality. Here, we use the inner product between $\boldsymbol{u}$ and $\boldsymbol{v}$ to measure the similarity of the two person images,

$$\begin{aligned} \texttt{Sim}(\boldsymbol{I}_u, \boldsymbol{I}_v) &= \Big\langle \frac{1}{|\mathcal{S}|} \sum_{p_u \in \mathcal{S}} \boldsymbol{U}_{p_u}, \frac{1}{|\mathcal{S}|} \sum_{p_v \in \mathcal{S}} \boldsymbol{V}_{p_v} \Big\rangle \\ &= \frac{1}{|\mathcal{S}|^2} \sum_{p_u, p_v \in \mathcal{S}} \left\langle \boldsymbol{U}_{p_u}, \boldsymbol{V}_{p_v} \right\rangle \end{aligned}, \tag{2}$$

where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle$ denotes the inner product between $\boldsymbol{u}$ and $\boldsymbol{v}$. The similarity of $\boldsymbol{u}$ and $\boldsymbol{v}$ can be interpreted as an average of part similarities between $|\mathcal{S}|^2$ part pairs.

However, such coarse similarity aggregation may degenerate into a suboptimal solution, which can be attributed to two major reasons. The first reason is associated with the imbalanced quantity distribution between about $|\mathcal{S}|$ aligned and $|\mathcal{S}|\,(|\mathcal{S}| - 1)$ misaligned body part pairs. Since the number of the misaligned pairs (shoulder $\leftrightarrow$ hand) is quadratically larger than the number of the aligned ones (hand $\leftrightarrow$ hand), the similarities of the aligned part pairs may be overwhelmed by the misaligned part pairs, which might exacerbate the part misalignment problem to some extent. The second reason is related to the non-person part descriptors containing various background clutters. This problem is particularly apparent when person bodies are partially occluded by other non-person objects. As a result, the background descriptors may bring an objectionable bias to the aggregated similarities in Eq. 2.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,

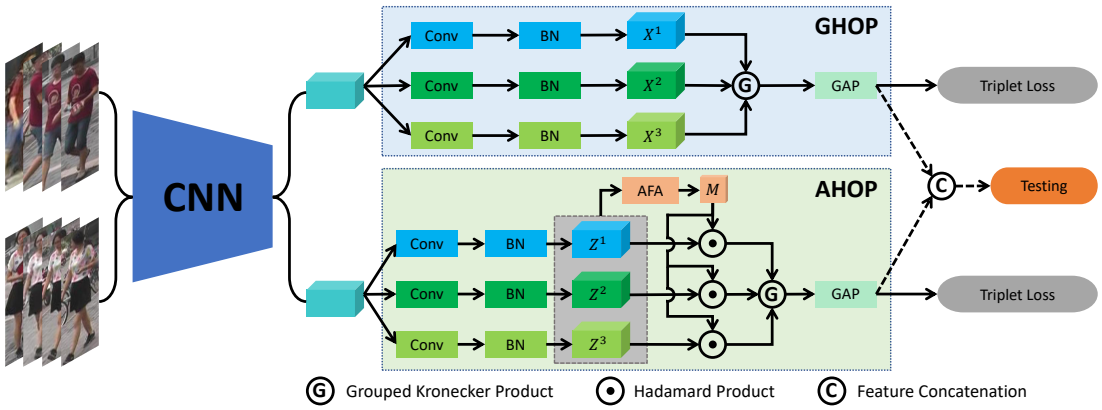


Fig. 3: Overview of the proposed GAReID framework. It consists of three parts, *i.e.*, a backbone network, a GHOP layer and an AHOP layer. The backbone network is input with person images to extract convolutional feature maps. Then we use a series of 1×1 convolutional layers and BatchNorm layers to produce multiple input feature maps, *i.e.*, $\{\boldsymbol{X}^i\}_{i=1}^n$ and $\{\boldsymbol{Z}^i\}_{i=1}^n$. Next, those feature maps are fed into the GHOP and AHOP layers to generate high-order image and foreground features, respectively. The output features are supervised by triplet loss during training, while we concatenate the two features to compute cosine similarities during testing.

B. Grouped High-Order Pooling

High-Order Representation: As illustrated in Fig. 1, the aligned parts usually contain identical semantics while the misaligned parts have dissimilar semantics, so the aligned part similarities are likely to be larger than the misaligned part similarities. However, recent works are unable to exploit this prior knowledge efficiently, so similarity discrepancies between aligned and misaligned part pairs may not be sharp and easy to distinguish. As indicated in Fig. 2, the high-order function $f(x) = x^n\,(n > 1, x \geq 0)$ contributes to enlarging the similarity discrepancies between aligned and misaligned body-part pairs. Note that we need to add a ReLU layer after input features to ensure the part similarity is always non-negative. By taking this high-order function into Eq. 2, a high-order similarity is defined as,

$$\mathtt{Sim}\left(\boldsymbol{I}_u, \boldsymbol{I}_v; n\right) = \frac{1}{|\mathcal{S}|^2} \sum_{p_u, p_v \in \mathcal{S}} \left\langle \boldsymbol{U}_{p_u}, \boldsymbol{V}_{p_v} \right\rangle^n, \tag{3}$$

where $\langle \boldsymbol{u}, \boldsymbol{v} \rangle^n$ represents the nth-order part similarity between parts $\boldsymbol{u}$ and $\boldsymbol{v}$. As the order n increases, the aligned part similarities will dominate the aggregated similarity in Eq. 3. Therefore, the high-order mapping function is beneficial to solve the part misalignment problem without the requirement of auxiliary landmark knowledges.

According to Theorem 1, the similarity of high-order features is equivalent to the high-order mapping of the first-order similarity. Subsequently, we reformulate Eq. 3 to simplify the computation of high-order similarities,

$$\begin{aligned} \mathtt{Sim}\left(\boldsymbol{I}_u, \boldsymbol{I}_v; n\right) &= \frac{1}{|\mathcal{S}|^2} \sum_{p_u, p_v \in \mathcal{S}} \left\langle \bigotimes_n \boldsymbol{U}_{p_u}, \bigotimes_n \boldsymbol{V}_{p_v} \right\rangle \\ &= \left\langle \frac{1}{|\mathcal{S}|} \sum_{p_u \in \mathcal{S}} \bigotimes_n \boldsymbol{U}_{p_u}, \frac{1}{|\mathcal{S}|} \sum_{p_v \in \mathcal{S}} \bigotimes_n \boldsymbol{V}_{p_v} \right\rangle \end{aligned}. \tag{4}$$

Hence, a high-order representation is defined as,

$$\boldsymbol{x} = \frac{1}{|\mathcal{S}|} \sum_{p_x \in \mathcal{S}} \bigotimes_n \boldsymbol{X}_{p_x}. \tag{5}$$

Since the Kronecker product $\otimes$ allows all elements of feature vectors to interact with each other, the high-order features exhibit strong representational capabilities. Notwithstanding, the dimension of high-order features increases exponentially, leading to very high memory consumption $\mathcal{O}\left(C^n\right)$ and computational complexity $\mathcal{O}\left(C^n\right)$. Therefore, an effective feature compression approach is needed to project high-order features onto a lower dimensional space.

High-Order Compression: Motivated by light-weight network design [24, 25], we propose a novel *Grouped Kronecker Product* (GKP) to compress high-order features by using channel group and shuffle strategies. As shown in Fig. 4, input feature channels are uniformly divided into G groups which are then shuffled to help the information dispersion across feature groups. Then, we employ the conventional Kronecker product for sub-features in each original and shuffled group, which contributes to encoding both intra-group and inter-group high-order interactions. Since the Kronecker product increases feature dimensions in each group, we further compress high-order features by conducting element-wise aggregation. This can significantly improve the effectiveness of high-order features. Mathematically, the nth-order GKP operation is formulated as,

$$\overset{G}{\underset{n}{\otimes}} \boldsymbol{x} = \begin{cases} \boldsymbol{x}, n = 1, \\ \sum\limits_{j=1}^{G} \boldsymbol{I}_j \boldsymbol{x} \otimes \boldsymbol{I}_j \boldsymbol{x} + \widehat{\boldsymbol{I}}_j \boldsymbol{x} \otimes \widehat{\boldsymbol{I}}_j \boldsymbol{x}, n = 2, \\ \sum\limits_{j=1}^{G} \left(\boldsymbol{I}_j \overset{G}{\underset{n-1}{\otimes}} \boldsymbol{x}\right) \otimes \boldsymbol{I}_j \boldsymbol{x} + \left(\widehat{\boldsymbol{I}}_j \overset{G}{\underset{n-1}{\otimes}} \boldsymbol{x}\right) \otimes \widehat{\boldsymbol{I}}_j \boldsymbol{x}, n > 2, \end{cases} \tag{6}$$

where $\boldsymbol{I}_j \in \mathbb{R}^{\frac{C}{G} \times C}$ is a block matrix and $\boldsymbol{I} = [\boldsymbol{I}_1; \boldsymbol{I}_2; \cdots; \boldsymbol{I}_G]$ is an identity matrix. $\widehat{\boldsymbol{I}} \in \mathbb{R}^{C \times C}$ is the shuffled version of the identity matrix $\boldsymbol{I}$. Note that we set $G = \sqrt{C}$ to keep high-order feature dimension unchanged. In this way, the proposed GKP has much lower time complexity $\mathcal{O}\left(nC^{2.5}\right)$ and space complexity $\mathcal{O}\left(nC^{2.5}\right)$ than the conventional Kronecker product. If $\sqrt{C}$ is not an integer, we set $G = \lceil \sqrt{C} \rceil$ and an extra sub-feature with a length $G^2 - C$ is generated by randomly sampling elements from the input feature. Then, we concatenate the input feature and sampled sub-feature to produce a new feature with length G^2. This fused feature is used to generate a high-order feature with length G^2. Then we randomly discard $G^2 - C$ elements from the high-order feature to reduce feature length to C.

High-Order Pooling: By applying the GKP into Eq. 5, the

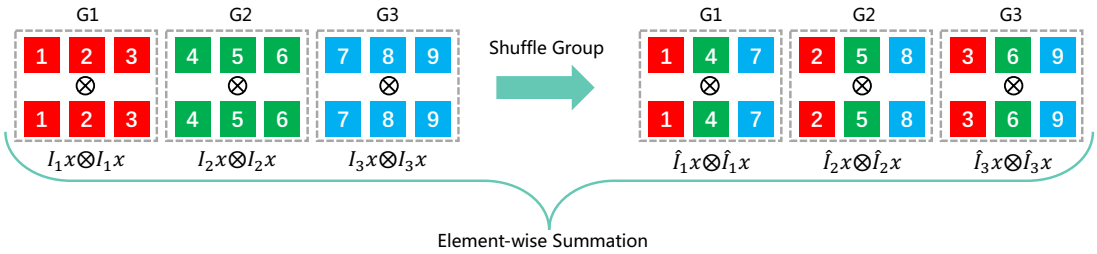


Fig. 4: A toy example of *Grouped Kronecker Product* (GKP) $\otimes_n^G \boldsymbol{x}$ with $n = 2$, $G = 3$ and $C = 9$. "G1", "G2" and "G3" represent the first, second and third group, respectively. The left two vectors are split into 3 groups and then the two subvectors of each group are aggregated by the conventional Kronecker product to produce a second-order vector with a length 9. The right two vectors are the group-shuffled versions of the left two vectors and then the three second-order vectors are generated by the same process as the left two vectors. Finally, all six second-order vectors are fused by the element-wise summation to produce a second-order vector with a length 9.

proposed GHOP layer is defined as,

$$\boldsymbol{x} = \frac{1}{|\mathcal{S}|} \sum_{p_x \in \mathcal{S}} \bigotimes_{n}^{G} \boldsymbol{X}_{p_x}. \tag{7}$$

Since multiple input features provide informative semantic characteristics of person poses, the high-order interactions among multiple input features are able to enhance the generalization ability of the GAReID model. For exploiting those high-order interactions, we extend the GHOP layer by reformulating Eq. 7 with multiple input features,

$$\boldsymbol{x} = \frac{1}{|\mathcal{S}|} \sum_{p_x \in \mathcal{S}} \boldsymbol{X}^1_{p_x} \bigotimes^{G} \boldsymbol{X}^2_{p_x} \bigotimes^{G} \cdots \bigotimes^{G} \boldsymbol{X}^n_{p_x}, \tag{8}$$

where $\otimes^G$ denotes the second-order GKP with $n = 2$. It is worth noting that this GHOP layer can be viewed as the high-order fusion method of multiple input features, which contributes to mining much richer information than the first-order method such as channel concatenation.

C. Attentive High-Order Pooling

Foreground Attention: Since aligned background similarities might introduce noise to the similarity aggregation of Eq. 2, the background knowledge should be excluded from person features. Recent studies [56] have found that the largest feature norms appear above target objects in a classification model pretrained on ImageNet. Our goal is to bootstrap on this phenomenon in order to highlight foreground regions without explicitly introducing learnable parameters. To this end, we design an attention module named *Adaptive Foreground Attention* (AFA) to produce a binary mask over spatial locations with using the l_2-norm of spatial features. Formally, given a feature map $\boldsymbol{Z} \in \mathbb{R}^{C \times H \times W}$, we first generate a feature map $\boldsymbol{T} \in \mathbb{R}^{H \times W}$ by operating the l_2 norm for features as,

$$\boldsymbol{T}_p = \left\| \boldsymbol{Z}_p \right\|_2, \tag{9}$$

where $\boldsymbol{T}_p$ denotes the response score of $\boldsymbol{T}$ at the position p. In order to mine foreground parts, we sample the positions where the response value is larger than an adaptive threshold. In this way, we produce a foreground position set as,

$$\mathcal{S}_F = \{p | \boldsymbol{T}_p > \varepsilon \boldsymbol{T}_{\text{avg}}\}, \tag{10}$$

where $\boldsymbol{T}_{\text{avg}}$ denotes the average response of $\boldsymbol{T}$ and $\varepsilon = 0.4$ is a hyperparameter controlling the activation threshold. Subsequently, the attention mask $\boldsymbol{M} \in \mathbb{R}^{H \times W}$ is formed as,

$$\boldsymbol{M}_p = \alpha \mathcal{I}\left(p \in \mathcal{S}_F\right) + \beta \mathcal{I}\left(p \notin \mathcal{S}_F\right), \ \forall p \in \mathcal{S}, \tag{11}$$

where $\boldsymbol{M}_p$ denotes the attention score of $\boldsymbol{M}$ at the position p. The indicator function $\mathcal{I}(\cdot)$ returns 1 if the input condition is true; otherwise it returns 0. In our experiments, we set foreground and background attention values as $\alpha = 1.0$ and $\beta = 0.3$, respectively.

Ensemble Attention: However, a single attention mask may not locate the foreground regions accurately because of diverse variations from person images. Inspired by ensemble learning, we adopt an element-wise average to fuse multiple attention masks generated from input feature maps in Eq. 8 as,

$$\boldsymbol{M} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}^i, \tag{12}$$

where $\boldsymbol{M}^i$ denotes the attention mask of the ith input feature map $\boldsymbol{Z}^i$. With the combination of the GHOP and AFA layers, the proposed AHOP layer is defined as follows,

$$\boldsymbol{z} = \frac{1}{|\mathcal{S}|} \sum_{p_z \in \mathcal{S}} \left(\boldsymbol{M}_{p_z} \boldsymbol{Z}^1_{p_z}\right) \bigotimes^{G} \cdots \bigotimes^{G} \left(\boldsymbol{M}_{p_z} \boldsymbol{Z}^n_{p_z}\right). \tag{13}$$

It is worth noting that this AHOP layer can be viewed as an attention fusion method, which aggregates different attention masks to refine the segmentation of foreground and background regions.

D. Overall Loss Function

In order to train the GAReID framework, we utilize the triplet loss [58] to learn discriminative high-order features. We define $\boldsymbol{x}_a$, $\boldsymbol{x}_p$ and $\boldsymbol{x}_n$ as the anchor, positive and negative high-order features from the GHOP layer, while $\boldsymbol{z}_a$, $\boldsymbol{z}_p$ and $\boldsymbol{z}_n$ represent the anchor, positive and negative high-order features from the AHOP layer. The triplet loss aims at separating the positive pair from the negative one by a similarity margin m. The triplet loss is defined as,

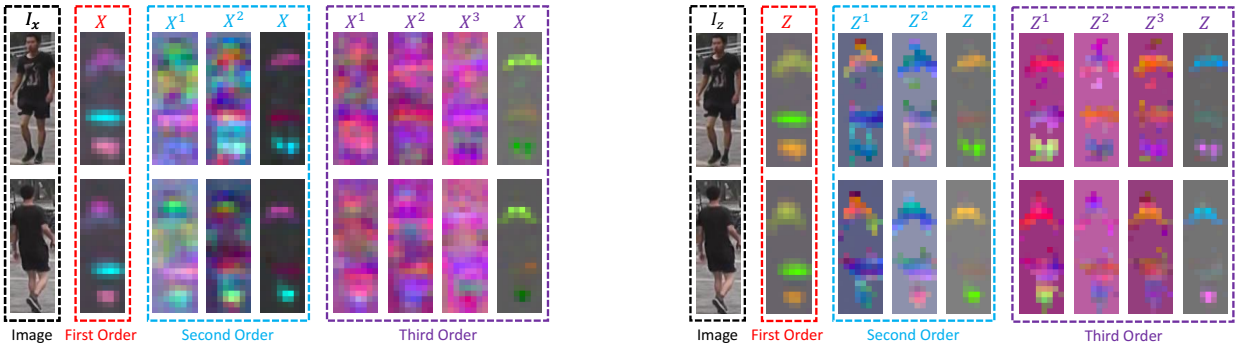
$$\begin{aligned} \mathcal{L}_t = \sum_{a,p,n} & \left[\left\langle \widetilde{\boldsymbol{x}}_a, \widetilde{\boldsymbol{x}}_n \right\rangle - \left\langle \widetilde{\boldsymbol{x}}_a, \widetilde{\boldsymbol{x}}_p \right\rangle + m\right]_+ \\ & + \left[\left\langle \widetilde{\boldsymbol{z}}_a, \widetilde{\boldsymbol{z}}_n \right\rangle - \left\langle \widetilde{\boldsymbol{z}}_a, \widetilde{\boldsymbol{z}}_p \right\rangle + m\right]_+ \end{aligned}, \tag{14}$$

where m is set as $m = 0.2$. The vectors $\widetilde{\boldsymbol{x}}_a$, $\widetilde{\boldsymbol{x}}_p$ and $\widetilde{\boldsymbol{x}}_n$ are the l_2 normalized features of $\boldsymbol{x}_a$, $\boldsymbol{x}_p$ and $\boldsymbol{x}_n$, while $\widetilde{\boldsymbol{z}}_a$, $\widetilde{\boldsymbol{z}}_p$ and $\widetilde{\boldsymbol{z}}_n$ are the l_2 normalized features of $\boldsymbol{z}_a$, $\boldsymbol{z}_p$ and $\boldsymbol{z}_n$.

IV. Discussion

A. Feature Visualization

In this part, considering the collaborative effect of high-order interactions among multiple feature maps in Eq. 8 and 13, we give a microscopic interpretation from the perspective of feature visualization, which shows a strong justification of our method. To some extent, it also reveals the reason why



(a) GHOP **(b)** AHOP

Fig. 5: Visualization of feature maps extracted from the first-order ($n = 1$), second-order ($n = 2$) and third-order ($n = 3$) GHOP and AHOP layers. Following the SIFTFlow [57], we use *Principal Component Analysis* (PCA) to compress all part descriptors into three-dimensional vectors and then rescale the vector values into the range of $[0, 255]$ to represent the three color channels of RGB images. In the visualized feature maps, the same color implies that the part descriptors are similar, whereas different colors indicate the part descriptors are dissimilar. Notably, $\{\boldsymbol{X}^i\}_{i=1}^n$ and $\{\boldsymbol{Z}^i\}_{i=1}^n$ represent the input feature maps of the nth-order GHOP and AHOP layers, while $\boldsymbol{X}$ and $\boldsymbol{Z}$ represent the output feature maps of the nth-order GHOP and AHOP layers.

the high-order interactions in the GHOP and AHOP layers contribute significantly to part-aligned and background-robust representation learning.

As exemplified in Fig. 5, one can observe that the first-order input feature maps mainly encode the semantics of various body parts, including heads, hands, shoulders and legs, and their corresponding colors differ depending on their spatial positions. Furthermore, the part descriptors with the same positions from different input feature maps are shown in different colors due to the diversity of multiple input feature maps. In Fig. 5a, the high-order output feature maps concentrate on encoding the discriminative body parts (*e.g.*, heads, shoulders and legs) to represent person identities, while the low-order output feature maps focus on capturing coarse-grained appearance information. Hence, the high-order interactions from the GHOP layer are beneficial as they enhance pose-invariance within the learned person features. In Fig. 5b, the proposed AHOP layer is able to remove background regions and retain foreground areas of the input feature maps. This contributes to high-order background-invariant representation learning.

B. Similarity Visualization

Based on the high-order similarity aggregation in Eq. 3, we provide another macroscopic explanation from high-order feature similarities. In a sense, it also furnishes a valuable angle for the understanding of the relationship between high-order feature similarities and fine-grained part alignments.

As illustrated in Fig. 6, the maximum part similarity of the high-order features is clearly larger than the similarity of the low-order features, while the minimum part similarity remains largely unchanged for all orders. In addition, the number of misaligned part pairs with prominent similarities consistently decreases along with the increase of feature order. Compared with the GHOP layer, the AHOP layer distinctly reduces the similarities of background part pairs, which reinforces the background-robust representation learning. Moreover, the increase amplitude of the maximum part similarity in the AHOP layer is evidently larger than the similarity of the GHOP layer with the same increase of feature orders. This observation indicates that the background removal alleviates the part misalignment problem.

C. Landmark Visualization

As suggested in prior works [8–10], the semantic knowledge of person landmarks is likely to remain unchanged, even when drastic pose variations have taken place. Besides, person pose variations mainly reflect the landmark distribution of person images. Therefore, to analyze the effectiveness of part alignments, it is worth exploring the high-order semantic interactions between different landmark pairs.

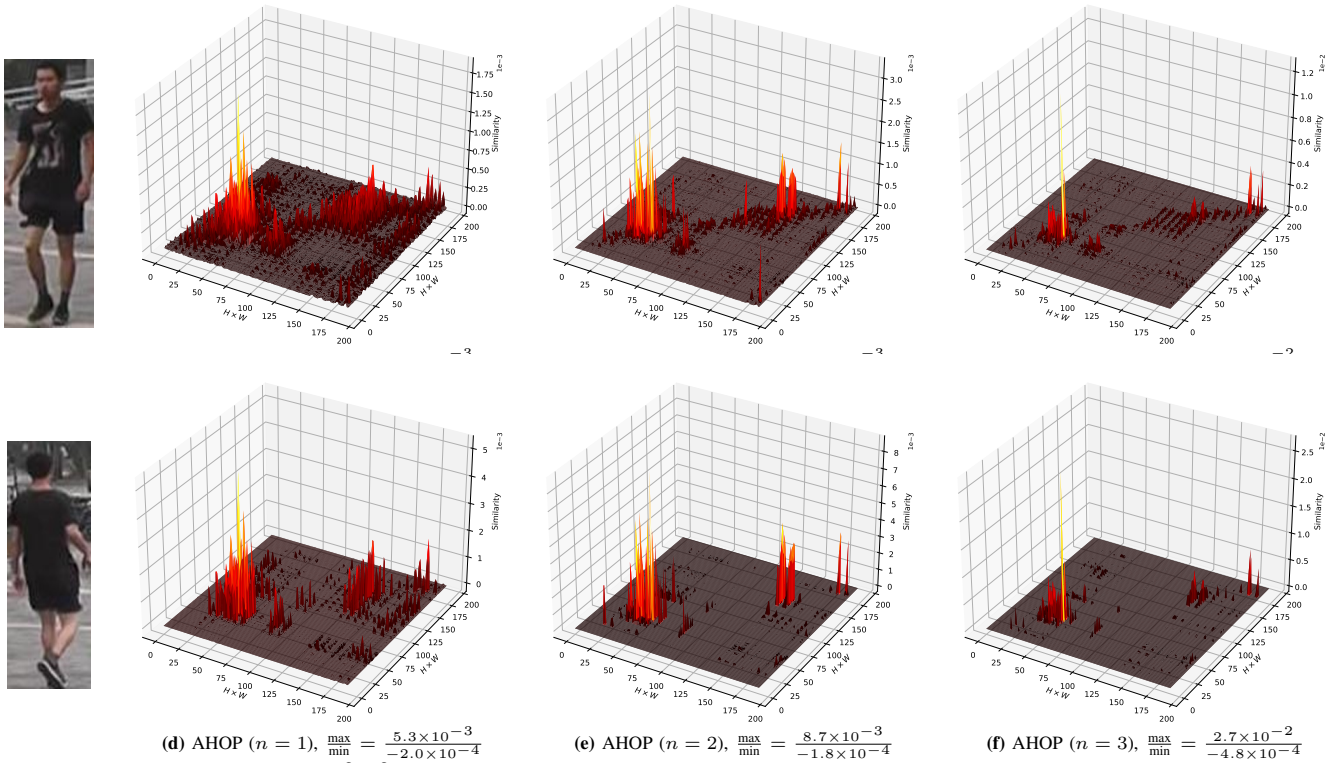
Given a pair of images $\boldsymbol{I}_u$ and $\boldsymbol{I}_v$ of the same person, we extract a pair of output feature maps $\boldsymbol{U}$ and $\boldsymbol{V}$ from the first-order, second-order, and third-order GHOP/AHOP layers. Then, we adopt an existing OpenPose [34] to detect 16 body landmarks for the two person images. In order to extract landmark descriptors, we upsample the output feature maps with the cubic interpolation to have the same size as the input images and then the landmark features are acquired from the resized feature maps according to landmark positions. Finally, the cosine similarities of 16×16 landmark pairs from the two images are computed to form a similarity confusion matrix. The results shown in Fig. 7 well demonstrate that the high-order part features can successfully learn landmark correspondences between the two images without using landmark annotations. Specifically, compared with the low-order features, the high-order features are able to significantly enlarge the similarity discrepancies between aligned and misaligned landmark pairs. More interestingly, the comparison between the GHOP and AHOP layer certifies that foreground region mining is conducive to highlighting the semantic correspondences of person landmarks.

D. Attention Visualization

In this part, we provide an intuitive interpretation by visualizing foreground attention masks to study the impact of the proposed AFA module. The visualized results demonstrate the superiority of the proposed ensemble attention strategy for the person ReID task. To some degree, the interpretation also clarifies the reason why learning foreground-based features is more helpful to part alignments than learning image-based features.

Given an input person image $\boldsymbol{I}_x$, we extract individual attention masks $\{\boldsymbol{M}^i\}_{i=1}^n$ and ensemble attention masks $\boldsymbol{M}$ from the first-order ($n = 1$), second-order ($n = 2$) and third-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,



(d) AHOP ($n=1$), $\frac{\max}{\min}=\frac{5.3\times10^{-3}}{-2.0\times10^{-4}}$ **(e)** AHOP ($n=2$), $\frac{\max}{\min}=\frac{8.7\times10^{-3}}{-1.8\times10^{-4}}$ **(f)** AHOP ($n=3$), $\frac{\max}{\min}=\frac{2.7\times10^{-2}}{-4.8\times10^{-4}}$

Fig. 6: Part similarity visualization of H^2W^2 part pairs. Given a pair of images from the same person, we extract a pair of output feature maps from the first-order ($n=1$), second-order ($n=2$) and third-order ($n=3$) GHOP/AHOP layers. The two feature maps are individually normalized by dividing the l_2 norms of spatially pooled features. Finally, the similarity matrix is calculated by the inner product of all H^2W^2 part pairs. Note that "max" and "min" denote the maximal and minimal part similarities, respectively.

order ($n=3$) AHOP layers. To better visualize the spatial relationships between confidence maps and body parts, the low-resolution attention mask is upsampled using the cubic interpolation to have the same size as $\boldsymbol{I}_x$. Then we merge both attention masks and person images by alpha blending. For interpreting the effectiveness of the proposed AFA method, we analyze the three attention generation methods, including "l_2 Norm", "Avg" and "Max". In particular, "l_2 Norm", "Avg" and "Max" represent that the l_2 norms, average values and maximal values along the channel dimension are used to generate attention masks, respectively. As seen in Fig. 8, "Avg" performs the worst among the three attention generation methods because it is unable to finely discriminate foreground regions from background ones. In other words, "Avg" mixes up foreground and background knowledge, which may hinder the background-robust representation learning. On the whole, both "l_2 Norm" and "Max" can successfully capture foreground regions and eliminate background areas without using person segmentation annotations. Compared with "Max", "l_2 Norm" performs foreground detection with a more fine-grained manner. For example, when the order factor $n=2$ or $n=3$, "Max" is unable to detect the foreground regions of person legs, while "l_2 Norm" is capable of avoiding a few residual background problems. Furthermore, as the order n increases, the foreground attention quality of "l_2 Norm" consistently improves with a significant margin. More interestingly, compared with individual attention masks $\{\boldsymbol{M}^i\}_{i=1}^n$, our ensemble attention mask $\boldsymbol{M}$ is beneficial to preserve discriminative foreground regions and remove hard background areas.

E. Similarity Attention

In order to analyze the impact of the proposed AHOP layer, we reformulate the high-order similarity between the two images $\boldsymbol{I}_u$ and $\boldsymbol{I}_v$ as,

$$\begin{aligned}&\texttt{Sim}\left(\boldsymbol{I}_u,\boldsymbol{I}_v;n\right)=\left\langle\boldsymbol{u},\boldsymbol{v}\right\rangle\\&=\frac{1}{|\mathcal{S}|^2}\sum_{p_u,p_v\in\mathcal{S}}\left\langle\boldsymbol{M}^u_{p_u}\boldsymbol{U}_{p_u},\boldsymbol{M}^v_{p_v}\boldsymbol{V}_{p_v}\right\rangle^n\\&=\frac{1}{|\mathcal{S}|^2}\sum_{p_u,p_v\in\mathcal{S}}\left(\boldsymbol{M}^u_{p_u}\boldsymbol{M}^v_{p_v}\right)^n\left\langle\boldsymbol{U}_{p_u},\boldsymbol{V}_{p_v}\right\rangle^n\end{aligned},\tag{15}$$

where $\boldsymbol{M}^u$ and $\boldsymbol{M}^v$ represent foreground attention maps of the two images $\boldsymbol{I}_u$ and $\boldsymbol{I}_v$, respecively. Additionally, $\left(\boldsymbol{M}^u_{p_u}\boldsymbol{M}^v_{p_v}\right)^n$ can be viewed as the nth-order similarity attention between p_u and p_v. To illustrate the effectiveness of foreground attention, we consider four part-pair cases, *i.e.*, *Foreground-Foreground* (FF), *Foreground-Background* (FB), *Background-Foreground* (BF) and *Background-Background* (BB). If the body-part pair belongs to the FF case, then $\boldsymbol{M}^u_{p_u}\boldsymbol{M}^v_{p_v}=1$ always holds and its high-order similarity attention keeps unchanged as follows,

$$\lim_{n\to\infty}\left(\boldsymbol{M}^u_{p_u}\boldsymbol{M}^v_{p_v}\right)^n=1.\tag{16}$$

If the part pair belongs to FB, BF or BB, then $\boldsymbol{M}^u_{p_u}\boldsymbol{M}^v_{p_v}<1$ always holds and its high-order similarity attention dramatically decreases as follows,

$$\lim_{n\to\infty}\left(\boldsymbol{M}^u_{p_u}\boldsymbol{M}^v_{p_v}\right)^n=0.\tag{17}$$

To sum up, the high-order similarity attention contributes to reducing the part similarity of FB, BF and BB pairs, while maintaining the similarity of FF pairs. As the order factor $n\to\infty$, the person similarity is equivalent to an average of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,

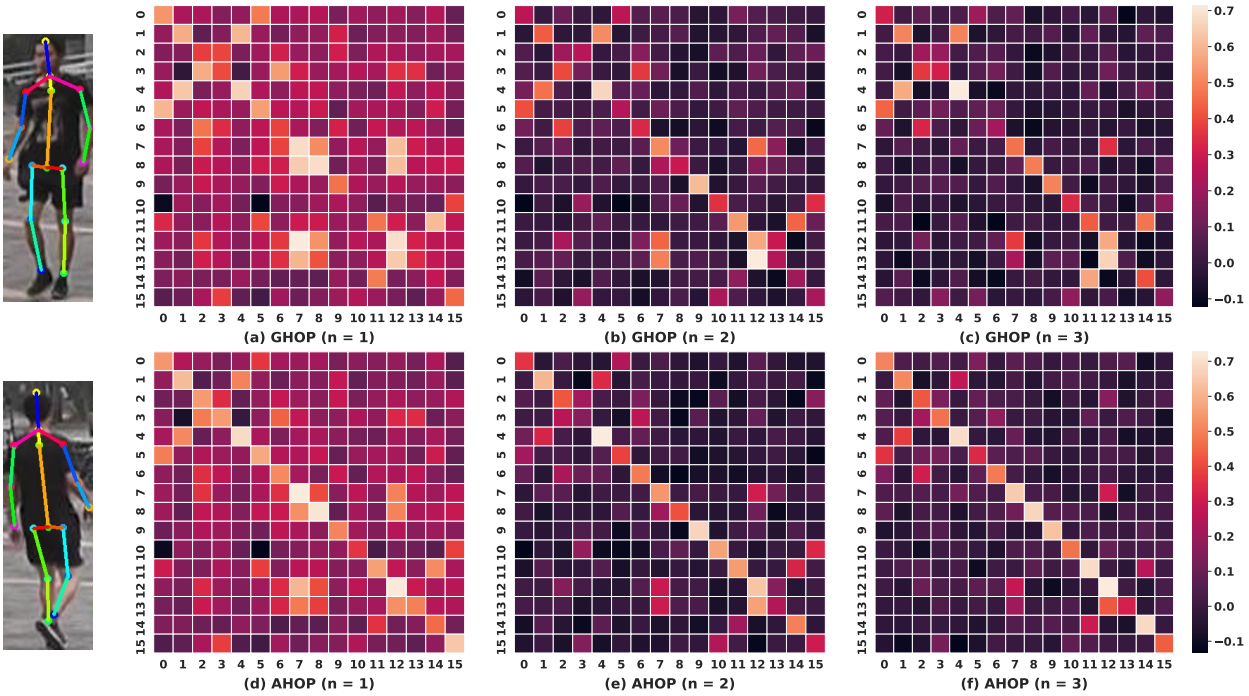


Fig. 7: Landmark similarity visualization of a pair of images from the same person. We employ OpenPose [34] to detect 16 body landmarks and then 16 landmark features are extracted from the first-order ($n = 1$), second-order ($n = 2$) and third-order ($n = 3$) GHOP/AHOP layers. Then, the cosine similarities of 16×16 landmark pairs from the two images are computed to form a similarity confusion matrix.

the similarities of aligned foreground part pairs, resulting in both part-aligned and background-robust person ReID.

F. Gradient Optimization

Finally, to assess the collaborative impact of high-order features on metric learning, we provide another theoretical analysis based on the gradient optimization for the triplet loss.

In order to simplify the following analysis, we ignore the l_2 normalization for high-order features. If we suppose that the high-order features are directly aggregated by the Kronecker product, the triplet loss is formulated as,

$$\begin{aligned}\mathcal{L} &= \Big[\big\langle \boldsymbol{z}_a, \boldsymbol{z}_n\big\rangle - \big\langle \boldsymbol{z}_a, \boldsymbol{z}_p\big\rangle + m_t\Big]_+ \\ &= \Big[\frac{1}{|\mathcal{S}|^2}\sum_{p_a,p_n\in\mathcal{S}}\big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^n_{p_n}\big)^{n_o}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^n_{p_n}\big\rangle^{n_o} \\ &\quad - \frac{1}{|\mathcal{S}|^2}\sum_{p_a,p_p\in\mathcal{S}}\big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^p_{p_p}\big)^{n_o}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^p_{p_p}\big\rangle^{n_o} + m_t\Big]_+ ,\end{aligned} \tag{18}$$

where n_o is the order coefficient of attentive high-order features. In addition, $\boldsymbol{Z}^a_{p_a}$ denotes the part feature vector of the anchor feature map $\boldsymbol{Z}^a$ at the position p_a, while $\boldsymbol{M}^a_{p_a}$ refers to the attention value of the anchor attention mask $\boldsymbol{M}^a$ at the position p_a. In the same way, a similar definition is also adopted to the positive/negative feature maps ($\boldsymbol{Z}^p_{p_p}$ and $\boldsymbol{Z}^n_{p_n}$) and attention masks ($\boldsymbol{M}^p_{p_p}$ and $\boldsymbol{M}^n_{p_n}$). For optimizing Eq. 18, we can calculate its gradient with respect to $\boldsymbol{Z}^a_{p_a}$, $\boldsymbol{Z}^p_{p_p}$ and $\boldsymbol{Z}^n_{p_n}$ as follows,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{Z}^a_{p_a}} &= \frac{n_o}{|\mathcal{S}|^2}\sum_{p_n\in\mathcal{S}}\big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^n_{p_n}\big)^{n_o-1}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^n_{p_n}\big\rangle^{n_o-1}\boldsymbol{Z}^n_{p_n} \\ &\quad - \frac{n_o}{|\mathcal{S}|^2}\sum_{p_p\in\mathcal{S}}\big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^p_{p_p}\big)^{n_o}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^p_{p_p}\big\rangle^{n_o-1}\boldsymbol{Z}^p_{p_p}, \\ \frac{\partial\mathcal{L}}{\partial\boldsymbol{Z}^p_{p_p}} &= -\frac{n_o}{|\mathcal{S}|^2}\sum_{p_a\in\mathcal{S}}\big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^p_{p_p}\big)^{n_o}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^p_{p_p}\big\rangle^{n_o-1}\boldsymbol{Z}^a_{p_a}, \\ \frac{\partial\mathcal{L}}{\partial\boldsymbol{Z}^n_{p_n}} &= \frac{n_o}{|\mathcal{S}|^2}\sum_{p_a\in\mathcal{S}}\big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^n_{p_n}\big)^{n_o}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^n_{p_n}\big\rangle^{n_o-1}\boldsymbol{Z}^a_{p_a},\end{aligned} \tag{19}$$

if the margin constraint of Eq. 18 is violated, or zero otherwise. To simplify the above formulas, we define

$$\begin{aligned}\boldsymbol{W}^{an}_{p_ap_n} &= \big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^n_{p_n}\big)^{n_o}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^n_{p_n}\big\rangle^{n_o-1}, \\ \boldsymbol{W}^{ap}_{p_ap_p} &= \big(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^p_{p_p}\big)^{n_o}\big\langle \boldsymbol{Z}^a_{p_a}, \boldsymbol{Z}^p_{p_p}\big\rangle^{n_o-1},\end{aligned} \tag{20}$$

where $\boldsymbol{W}^{an}_{p_ap_n}$ and $\boldsymbol{W}^{ap}_{p_ap_p}$ can be viewed as two weight coefficients for different part pairs. Consequently, Eq. 19 can be rewritten as,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\boldsymbol{Z}^a_{p_a}} &= \frac{n_o}{|\mathcal{S}|^2}\sum_{p_n\in\mathcal{S}}\boldsymbol{W}^{an}_{p_ap_n}\boldsymbol{Z}^n_{p_n} - \frac{n_o}{|\mathcal{S}|^2}\sum_{p_p\in\mathcal{S}}\boldsymbol{W}^{ap}_{p_ap_p}\boldsymbol{Z}^p_{p_p}, \\ \frac{\partial\mathcal{L}}{\partial\boldsymbol{Z}^p_{p_p}} &= -\frac{n_o}{|\mathcal{S}|^2}\sum_{p_a\in\mathcal{S}}\boldsymbol{W}^{ap}_{p_ap_p}\boldsymbol{Z}^a_{p_a}, \\ \frac{\partial\mathcal{L}}{\partial\boldsymbol{Z}^n_{p_n}} &= \frac{n_o}{|\mathcal{S}|^2}\sum_{p_a\in\mathcal{S}}\boldsymbol{W}^{an}_{p_ap_n}\boldsymbol{Z}^a_{p_a}.\end{aligned} \tag{21}$$

According to Eq. 21, the gradient term with respect to $\boldsymbol{Z}^a_{p_a}$ is equivalent to the difference between the weighted average of all positive and negative part descriptors. Similarly, the gradient terms with respect to $\boldsymbol{Z}^p_{p_p}$ and $\boldsymbol{Z}^n_{p_n}$ are equivalent to the weight average of all part descriptors of anchor samples.

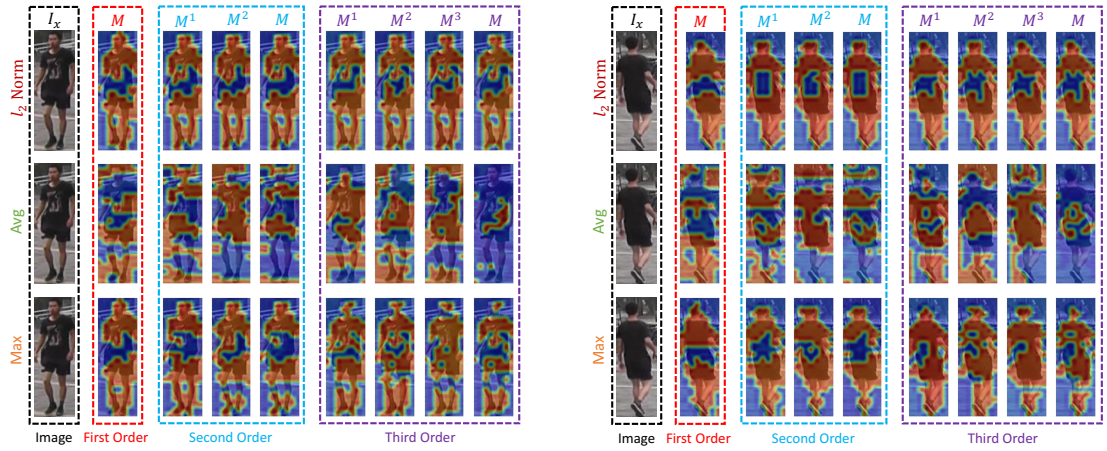


Fig. 8: Attention mask visualization of the three attention methods. In each visualized attention, the red areas refer to the foreground regions while the purple ones are associated with the background clutters. We extract individual attention masks $\{\boldsymbol{M}^i\}_{i=1}^n$ and ensemble attention masks $\boldsymbol{M}$ from the first-order ($n = 1$), second-order ($n = 2$) and third-order ($n = 3$) AHOP layers. "l_2 Norm", "Avg" and "Max" represent that the l_2 norms, average values and maximal values along the channel dimension are used to generate attention masks, respectively.

When $n_o \geq 2$, the two weight coefficients $\boldsymbol{W}^{an}_{p_a p_n}$ and $\boldsymbol{W}^{ap}_{p_a p_p}$ can be viewed as attentive (n_o-1)th-order feature similarities. As the order n_o increases, the gradients of aligned part descriptors are highlighted over those of the misaligned parts. In this case, the gradient term $\partial\mathcal{L}/\partial\boldsymbol{Z}^a_{p_a}$ pushes the anchor descriptor $\boldsymbol{Z}^a_{p_a}$ close to the aligned positive part descriptors and away from the aligned negative part descriptors. Likewise, $\partial\mathcal{L}/\partial\boldsymbol{Z}^p_{p_p}$ pushes the positive part descriptor $\boldsymbol{Z}^p_{p_p}$ close to the aligned anchor part descriptors, whilst $\partial\mathcal{L}/\partial\boldsymbol{Z}^n_{p_n}$ keeps the negative part descriptor $\boldsymbol{Z}^n_{p_n}$ away from the aligned anchor part descriptors. On the basis of the analysis in Sec. III-C, the attention factors, *i.e.*, $\left(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^n_{p_n}\right)^{n_o}$ and $\left(\boldsymbol{M}^a_{p_a}\boldsymbol{M}^p_{p_p}\right)^{n_o}$, are able to reduce the background effects on similarity aggregation. Accordingly, the composition of background gradients is effectively eliminated for Eq. 21. To some degree, this explains why the background removal contributes to enabling part alignment for the ReID task.

When $n_o = 1$, the two weight coefficients are reformulated as $\boldsymbol{W}^{an}_{p_a p_n} = \boldsymbol{M}^a_{p_a}\boldsymbol{M}^n_{p_n}$ and $\boldsymbol{W}^{ap}_{p_a p_p} = \boldsymbol{M}^a_{p_a}\boldsymbol{M}^p_{p_p}$, respectively. Therefore, the weight coefficient of the first-order AHOP layer is equivalent to the product of two attention values. Although the attention product encodes the relationships between a pair of spatial positions, there is no guarantee that the attention product of the aligned part pair will be larger than the misaligned part. For example, if both hands and legs have very large attention values, the aligned (leg $\leftrightarrow$ leg and hand $\leftrightarrow$ hand) and misaligned (leg $\leftrightarrow$ hand) part pairs might have similar values of attention product. Thus, the gradient terms of the first-order model contain both aligned and misaligned part descriptors, which might generate even totally erroneous gradient directions for backpropagation optimization.

According to the above analysis, the weight coefficient can be treated as a regularization term to regulate the gradient direction, which explains well the reason why the high-order features enhance the generalization ability of the ReID model. In summary, by considering the collaborative effort of all gradient terms, we could understand better the working principle of the proposed GAReID framework. Our framework not only enables regularization for the gradient direction but also enhances the part-alignment and background-robustness properties within features.

V. Experiments

A. Dataset

Market1501 [26]: It contains 32,668 images of 1,501 persons captured by six camera views. The whole dataset is divided into a training set containing 12,936 images of 751 persons and a testing set containing 19,732 images of 750 persons. For each person in the testing set, we select one image from each camera as a query image, forming 3,368 queries following the standard setting in [26].

CUHK03 [27]: It contains 14,097 images of 1,467 persons, captured by six camera views. Two types of person images are provided: manually-labeled person bounding boxes (Labeled) and automatically-detected bounding boxes (Detected). We use the settings of both labeled and detected person images on the splits in [71], where 767 and 700 persons are used for training and testing, respectively.

DukeMTMC [28]: It contains 36,411 images of 1,812 persons captured by 8 cameras, where only 1,404 persons appeared in more than 2 cameras. The other 408 persons are regarded as distractors. The training set contains 16,522 images of 702 persons while the testing set contains 2,228 query images of 702 persons and 17,661 gallery images.

MSMT17 [29]: It contains manually annotated 126,441 bounding boxes of 4,101 persons, which is currently the largest person ReID dataset. All images are captured by the 15-camera network deployed in campus, which contains 12 outdoor cameras and 3 indoor cameras. The training set contains 32,621 bounding boxes of 1,041 persons, and the testing set contains 93,820 bounding boxes of 3,060 persons. From the testing set, 11,659 bounding boxes are randomly selected as query images and the other 82,161 bounding boxes are used as gallery images.

B. Implementation Details

Network Architecture: We take the ResNet-50/101 [59] initialized with the parameters pretrained on ImageNet [72] as the backbone network. Following the work [14], the last fully-connected layer and global average pooling layer are removed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,

TABLE I: Comparison with state-of-the-art methods on Market1501 [26], CUHK03 [27], DukeMTMC [28] and MSMT17 [29] datasets. CUHK03-L and CUHK03-D use labeled and detected bounding boxes to crop person images on CUHK03, respectively. Two baseline models based on ResNet50/101 [59] are trained with triplet loss and global features are extracted from the GAP layer to perform ReID evaluation.

Method	Market1501		CUHK03-L		CUHK03-D		DukeMTMC		MSMT17	
	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP	Rank1	mAP
PDC [60]	84.14	63.41	-	-	-	-	-	-	58.00	29.70
GLAD [1]	89.90	73.90	-	-	-	-	-	-	61.40	34.00
HACNN [50]	91.20	75.70	44.40	41.00	41.70	38.60	80.50	63.80	-	-
PAB [3]	91.70	79.60	-	-	-	-	84.40	69.30	-	-
PCB + RPP [2]	93.80	81.60	63.70	57.50	-	-	83.30	69.20	68.20	40.40
MGN [5]	95.70	86.90	68.00	67.40	66.80	66.00	88.70	78.40	-	-
IANet [61]	94.40	83.10	-	-	-	-	83.10	73.40	75.50	46.80
DSAReID [11]	95.70	87.60	78.90	75.20	78.20	73.10	86.20	74.30	-	-
MHN [42]	95.10	85.00	77.20	72.40	71.70	65.40	89.10	77.20	-	-
OSNet [62]	94.80	84.90	-	-	72.30	67.80	88.60	73.50	78.70	52.90
SAN [63]	96.10	88.00	80.10	76.40	79.40	74.60	87.90	75.50	79.20	55.70
RGA-SC [64]	96.10	88.40	81.10	77.40	79.60	74.50	-	-	80.30	57.50
LEAP-CF [65]	93.50	84.20	-	-	-	-	87.80	74.20	76.70	50.80
GASM [66]	95.30	84.70	-	-	-	-	88.30	74.40	79.50	52.50
PISNet [67]	95.60	87.10	-	-	-	-	88.80	78.70	-	-
ISP [68]	95.30	88.60	76.50	74.10	75.20	71.40	89.60	80.00	-	-
ReID-NAS [69]	95.10	85.70	-	-	-	-	88.10	74.60	79.50	53.30
Occluded-ReID [36]	92.70	81.30	-	-	-	-	86.20	72.60		
RFC [70]	95.20	89.20	-	-	81.10	78.00	90.70	80.70	82.00	60.20
ResNet50 + GAP	93.53	84.83	76.57	74.37	75.71	72.10	85.59	72.84	69.31	44.43
ResNet50 + GAReID	96.13	89.42	82.04	81.02	80.20	78.76	89.28	81.64	80.57	58.73
ResNet101 + GAP	93.82	86.58	81.21	78.84	77.21	75.13	86.94	76.01	73.63	49.75
ResNet101 + GAReID	**96.45**	**90.76**	**86.21**	**83.80**	**84.64**	**82.03**	**91.04**	**82.13**	**82.40**	**61.22**

and the stride of the last residual block *Conv4_1* is set from 2 to 1 for increasing the feature map size.

Data Processing: In order to obtain enough context information from person images and a proper size of feature map for the proposed LTReID framework, we first resize training images to 384×128. Then we randomly crop each training image with scale in the interval $[0.64, 1.0]$ and aspect ratio $[2, 3]$. Third, we resize these cropped images back to 384×128. Following the work [51], the training images are augmented with horizontal flipping and random erasing [73]. Before it is sent to the network, each image is subtracted from the mean values $[0.485, 0.456, 0.406]$ and divided by the standard deviations $[0.229, 0.224, 0.225]$ according to normalization procedure when using the pretrained model on ImageNet.

Training/Testing Configurations: Since triplet loss is used to learn person features, we need to adopt an appropriate triplet sampling strategy. To simplify this procedure, triplets are generated using the $\mathcal{PK}$ sampling method [74], which randomly samples $\mathcal{P}$ classes and then randomly selects $\mathcal{K}$ images for each person to form a mini-batch with the size $\mathcal{P} \times \mathcal{K}$. In a mini-batch, we use all possible $\mathcal{PK}\left(\mathcal{PK}-\mathcal{K}\right)\left(\mathcal{K}-1\right)$ combinations of triplets for triplet loss. For all datasets, $\mathcal{P}$ and $\mathcal{K}$ are set to 16 and 4, respectively. Following the work [3], we use the *Stochastic Gradient Descent* (SGD) algorithm to minimize the overall loss function, where the initial learning rate, weight decay and momentum are set to 0.01, 2×10^{-4} and 0.9, respectively. The learning rate is decreased by a factor of 5 after every 200 epochs and all models are trained for 750 epochs. As for the testing phase, we use the cosine distance to measure the similarities between the probe and gallery images. Besides, *Mean Average Precision* (mAP) and *Rank1* (R1) accuracy are used for evaluation. All our methods are implemented on PyTorch [75]. All experiments run on a server with 2 Intel(R) Xeon(R) E5-2620 v4@2.10GHz CPUs, 4 GeForce GTX 1080 Ti GPUs and 128G RAM.

TABLE II: Comparative experiments using different attention mechanism methods. "l_2 Norm", "Avg" and "Max" represent that the l_2 norms, average values and maximal values along the channel dimension are used to generate attention masks, respectively. Note that all models use ResNet50 as the backbone.

Method	mAP			
	Market1501	CUHK03-D	DukeMTMC	MSMT17
SE [76]	89.13	78.50	81.44	58.34
AG [62]	89.40	78.69	81.47	58.72
w/o. Atten	89.00	78.44	81.42	58.03
Avg	87.47	74.37	79.65	57.42
Max	88.33	76.95	80.12	57.79
l_2 Norm	**89.42**	**78.76**	**81.64**	**58.73**

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

In Table I, we compare the proposed GAReID with current state-of-the-art methods on the four person ReID datasets. From the results we can see that our method achieves the best ReID performance on each dataset. Specifically, the proposed GAReID based on ResNet50 outperforms the previous best performed SAN [63] by **4.16%** in mAP on CUHK03-D. Although our method performs closely to ISP [68] on Market1501 and DukeMTMC datasets, our method can achieve a slightly higher accuracy in a very simple yet effective way. This is because the GAReID performs superior part alignment with only identity labels while other methods require landmark annotations or body partition during the training and testing phases. Compared with other datasets, the MSMT17 dataset presents the following challenges: (1) large number of person identities, bounding boxes and cameras; (2) complex scenes and backgrounds; (3) multiple time slots with severe lighting changes. Although all compared methods achieve lower accuracies on MSMT17 than other datasets, the proposed GAReID is the best-performing method, outperforming the second best method by **1.23%** for mAP. This clearly demonstrates that the GAReID achieves a satisfactory generalization on the large-scale dataset.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,

TABLE III: Ablation studies of different modules on Market1501, CUHK03, DukeMTMC and MSMT17 datasets. "HOP", "MF", "GS" and "EA" represent high-order pooling, multiple feature input, group shuffle and ensemble attention, respectively. Note that all models use ResNet50 as the backbone.

Method				mAP			
HOP	MF	GS	EA	Market1501	CUHK03-D	DukeMTMC	MSMT17
✘	✘	✘	✘	84.83	72.10	72.84	44.43
✔	✘	✘	✘	87.69	76.01	78.83	51.46
✔	✔	✘	✘	88.52	77.63	80.50	56.99
✔	✔	✔	✘	89.00	78.44	81.42	58.03
✔	✔	✔	✔	**89.42**	**78.76**	**81.64**	**58.73**

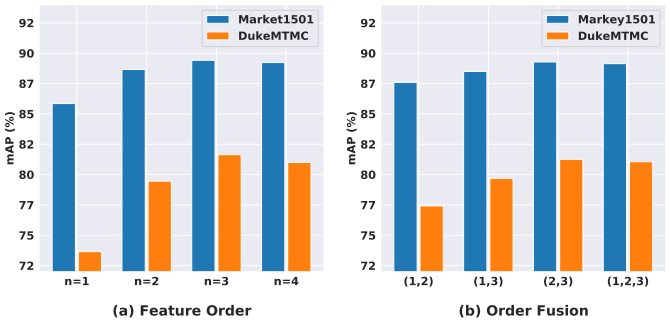


Fig. 9: Ablation studies on Market1501 [26] and DukeMTMC [28] datasets. (a) Analyzing the impact of the order n. (b) Comparing different order fusion strategies, "(1,2)" means that the first-order and second-order features are fused by channel concatenation.

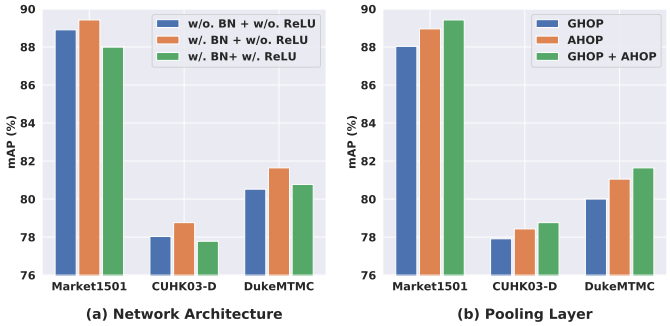


Fig. 10: Ablation studies on Market1501, CUHK03 and DukeMTMC datasets. (a) Analyzing different network architectures. (b) Analyzing different pooling layers.

D. Ablation Study

Feature Order: We first study the impact of the order of high-order features. As seen in Fig. 9(a), we can observe two interesting phenomena. First, a higher feature order benefits person ReID performance. The mAP scores of Market1501 and DukeMTMC datasets increase consistently until they reach a stable performance. For example, the third-order feature ($n = 3$) outperforms the first-order feature ($n = 1$) by **3.57%** and **7.99%** in terms of mAP on Market1501 and DukeMTMC datasets, respectively. Second, increasing the order ($n > 3$) makes a limited contribution to the mAP improvement compared with $n = 3$. To some extent, this is because the third-order pooling layer has largely eliminated part misalignments. Therefore, there is little room for further part alignment improvements. To sum up, we recommend $n = 3$ for the GAReID as it strikes a satisfactory balance between the computational efficiency and ReID performance.

Order Fusion: We explore the effectiveness of order fusion by averaging features from different orders. Two interesting observations can be made in Fig. 9(b). First, compared with low-order features ($n = (1, 2)$), fusing high-order features ($n = (2, 3)$) always benefits person ReID performance. The main reason is that the high-order features help to reduce the person part misalignment problem. Second, compared with single-order features ($n = 3$), mixed-order features ($n = (1, 2, 3)$) may significantly degrade ReID accuracies. To some extent, this is because fusing too many low-order feature is unable to highlight the discriminative information.

Attention Generation: We compare the performance of different attention generation methods on Market1501 and DukeMTMC datasets. The results in Table II show that the "l_2 Norm" consistently achieves superior mAP scores than other attention methods. This suggests that "l_2 Norm" is more suitable to mine foreground regions than other methods. Moreover, we observe that "Avg" achieves the worst mAP score than the model without using attention. As illustrated in Fig. 8, the main reason is that "Avg" can be viewed as a low-pass filter which removes some discriminative information. In view of performance and efficiency, we adopt the "l_2 Norm" to generate foreground attention masks in this work.

Multiple Feature Fusion: In this part, we examine the effectiveness of multiple feature fusion in Eq. 8 and 13. Specifically, the multiple feature fusion represents that multiple features are aggregated by the Kronecker product, while the single feature input denotes that the multiple duplicates of the single feature are aggregated by the Kronecker product. From the results in Table III, it can be observed that the multiple feature fusion performs better than the single feature input on the three datasets. The major reason is that the multiple features are able to bring richer pose knowledges than the single feature, resulting in a very strong high-order representational capability for the ReID models.

Group Shuffle: Since the channel group strategy is crucial to high-order feature compression, we need to explore the impact of the group shuffle strategy on enhancing the generalization capability of ReID models. From Table III, we can observe that the group shuffle strategy consistently improves ReID performance with a significant margin on the three datasets. This is because the group shuffle strategy encodes the inter-group interactions, which are beneficial to enrich the information of compressed high-order features.

Ensemble Attention: In this part, we investigate the impact of the ensemble attention on background-robust feature learning. We also design the independent attention masks for different input features to eliminate background regions. From the results reported in Table III, we notice that the ensemble attention achieves significant ReID performance improvements over the independent attention. This observation indicates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI10.1109/TNNLS.2022.3209537,

TABLE IV: Analysis of computational costs. "ResNet50 + Cat + GAP" represents multiple output features are concatenated (Cat) for testing, which can be viewed as the additional baseline model for "ResNet-50 + GHOP". "ResNet50 + Atten + Cat + GAP" denotes that foreground-based features are obtained by attention maps and then both image-based and foreground-based features are concatenated for testing, which can be viewed as the additional baseline model for "ResNet50 + GHOP + AHOP". The inference time and speed of all models are tested on a single GeForce GTX 1080 Ti GPU.

Method	#Features	#Parameters	FLOPS	Speed (fps)	Time (ms)	mAP			
						Market1501	CUHK03-D	DukeMTMC	MSMT17
ResNet50 + GAP	1	24.56M	6.28G	125.34	7.98	84.83	72.10	72.84	44.43
ResNet50 + Cat + GAP	2	25.61M	6.39G	122.76	8.15	85.27	72.87	73.01	45.76
ResNet50 + Atten + Cat + GAP	2	26.66M	6.50G	121.83	8.21	85.76	72.99	73.40	45.82
ResNet50 + CBP [43]	2	26.13M	6.43G	120.97	8.27	87.02	76.32	77.62	55.22
ResNet50 + MFB [77]	2	25.61M	6.39G	121.54	8.23	86.34	74.71	76.50	54.81
ResNet50 + DBT [40]	2	25.61M	6.40G	119.98	8.33	86.96	75.64	77.03	55.01
ResNet50 + GHOP	2	25.61M	6.40G	121.20	8.25	87.27	77.00	78.82	56.34
ResNet50 + GHOP + AHOP	2	26.66M	6.54G	119.87	8.34	88.66	77.34	79.46	57.26
ResNet50 + Cat + GAP	3	26.13M	6.46G	114.65	8.72	85.34	72.66	73.15	45.92
ResNet50 + Atten + Cat + GAP	3	27.71M	6.62G	110.07	9.09	85.57	72.90	73.36	46.27
ResNet50 + KP [39]	3	26.92M	6.53G	109.03	9.17	87.56	77.57	79.43	57.76
ResNet50 + MFH [78]	3	26.13M	6.46G	113.41	8.82	87.02	76.34	79.14	57.55
ResNet50 + GHOP	3	26.13M	6.50G	112.38	8.90	88.03	78.25	80.00	57.73
ResNet50 + GHOP + AHOP	3	27.71M	6.74G	109.88	9.10	**89.42**	**78.76**	**81.64**	**58.73**

that the ensemble attention, by integrating multiple attention masks, can reduce the influence of background clutters more effectively and generate better background-robust features.

Network Architecture: As shown in Fig. 3, we use different BN layers for different input branches. The impact of the BN layer for high-order feature learning can be observed in Fig. 10(a). As seen, using the BN layer achieves significant improvements on the three datasets, which justifies that the normalized features contribute to learning discriminative high-order representations. As mentioned in Sec. III-B, we need to add a ReLU layer after input features to ensure the part similarity is always non-negative. Therefore, it is worthwhile to examine whether the GAReID can perform satisfactory part alignments without ReLU. From the results in Fig. 10(a), we observe that the GAReID without ReLU achieves superior performance than the architecture with ReLU. This is because ReLU causes "dead" neurons when their activation values are negative. In other words, ReLU restricts the distribution of feature maps to a non-negative space and ignores the information of negative neurons. This might impair the representational capability of high-order features. In line with the findings of the above analysis, we use the BN layer instead of the ReLU layer as default.

Pooling Layer: In this part, we investigate the contributions of the GHOP and AHOP layers on part-aligned representation learning. In Fig. 10(b), the results show that the AHOP layer consistently achieves superior mAP scores than the GHOP layer. This phenomenon indicates that foreground-based features are more suitable than image-based features in part alignment tasks. To show the effectiveness of the joint learning of the two features, we concatenate them along the feature dimension to obtain complete person representations. We observe that the leveraging of these two features significantly outperforms either of them on the three datasets. In other words, the foreground-based features have some advantages over the image-based ones, but they are complementary to each other. Therefore, our proposed system simultaneously learns both image-based and foreground-based features.

Computational Costs: In Table IV, we further analyze the computational costs of the proposed GAReID framework. Compared with "ResNet50 + GAP", "ResNet50 + GHOP" achieves superior performances and only brings 1M $\sim$ 2M additional network parameters. Compared with "ResNet50 + Cat + GAP", "ResNet50 + GHOP" has similar computational costs but achieves higher mAP scores, which shows that the improvements of the GHP layers are not brought by the additional computational costs. Besides, "ResNet50 + GHOP + AHOP" performs better than "ResNet50 + Atten + Cat + GAP" when similar model parameters are used. We also compare the GAReID with other pooling methods, *i.e.*, CBP [43], MFB [77], DBT [40], KP [39] and MFH [78]. The comparison results show that the proposed GHOP and AHOP are able to encode more discriminative high-order features than other pooling methods. More interestingly, the running speed of the GAReID is more than 109 FPS, which is fast enough for video-based person ReID in real-time.

VI. Conclusion

In this paper, we propose a *Grouped Attentive Re-Identification* (GAReID) framework to alleviate the pose misalignment problem for person re-identification. The proposed framework designs the *Grouped High-Order Pooling* (GHOP) and *Attentive High-Order Pooling* (AHOP) layers to learn image-based and foreground-based high-order features, respectively. Besides, we put forward a novel feature compression method named *Grouped Kronecker Product* (GKP) to reduce the dimension of high-order features. Our theoretical analysis shows that high-order features facilitate pose alignments without depending on landmark detection or feature partition. Extensive experiments demonstrate that the GAReID framework achieves state-of-the-art performance on various person datasets. In the future, we will extend this work to the fields of attribute recognition, face recognition and vehicle re-identification, where the part misalignment problem is prevalent.

References

[1] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "Glad: Global-local-alignment descriptor for pedestrian retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 420–428.

[2] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 480–496.

[3] 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.

[4] L. Wu, Y. Wang, J. Gao, and X. Li, "Where-and-when to look: Deep siamese attention networks for video-based person re-identification," in *IEEE Transactions on Multimedia*, vol. 21, no. 6. IEEE, 2018, pp. 1412–1424.

[5] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 274–282.

[6] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 650–667.

[7] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," in *Advances in Neural Information Processing Systems*, 2018, pp. 1230–1241.

[8] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1179–1188.

[9] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2119–2128.

[10] M. M. Kalayeh, E. Basaran, M. Gökmen, M. E. Kamasak, and M. Shah, "Human semantic parsing for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1062–1071.

[11] Z. Zhang, C. Lan, W. Zeng, and Z. Chen, "Densely semantically aligned person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 667–676.

[12] 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*, vol. 43, no. 3, pp. 902–917, 2019.

[13] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8514–8522.

[14] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, "Horizontal pyramid matching for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8295–8302.

[15] P. Wang, Z. Zhao, F. Su, X. Zu, and N. V. Boulgouris, "Horeid: deep high-order mapping enhances pose alignment for person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 2908–2922, 2021.

[16] P. Wang, Z. Zhao, F. Su, and H. Meng, "Ltreid: Factorizable feature generation with independent components for long-tailed person re-identification," *IEEE Transactions on Multimedia*, 2022.

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

[18] 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.

[19] I. Kviatkovsky, A. Adam, and E. Rivlin, "Color invariants for person reidentification," in *IEEE Transactions on pattern analysis and machine intelligence*, vol. 35, no. 7. IEEE, 2012, pp. 1622–1634.

[20] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 144–151.

[21] Z. Wu, Y. Li, and R. J. Radke, "Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features," in *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5. IEEE, 2014, pp. 1095–1108.

[22] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptors with application to person re-identification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 9, pp. 2179–2194, 2019.

[23] J. Li, S. Zhang, Q. Tian, M. Wang, and W. Gao, "Pose-guided representation learning for person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 2, pp. 622–635, 2022.

[24] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[25] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[26] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[27] 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.

[28] 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.

[29] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 79–88.

[30] P. Wang, F. Su, Z. Zhao, Y. Zhao, L. Yang, and Y. Li, "Deep hard modality alignment for visible thermal person re-identification," *Pattern Recognition Letters*, vol. 133, pp. 195–201, 2020.

[31] P. Wang, Z. Zhao, F. Su, Y. Zhao, H. Wang, L. Yang, and Y. Li, "Deep multi-patch matching network for visible thermal person re-identification," *IEEE Transactions on Multimedia*, vol. 23, pp. 1474–1488, 2020.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[33] 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.

[34] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7291–7299.

[35] R. Alp Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7297–7306.

[36] J. Miao, Y. Wu, and Y. Yang, "Identifying visible parts via pose estimation for occluded person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 9, pp. 4624–4634, 2022.

[37] Y. Ding, H. Fan, M. Xu, and Y. Yang, "Adaptive exploration for unsupervised person re-identification," *ACM Transactions on Multimedia Computing, Communications, and Applications*

(TOMM), vol. 16, no. 1, pp. 1–19, 2020.
[38] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.
[39] Y. Cui, F. Zhou, J. Wang, X. Liu, Y. Lin, and S. Belongie, "Kernel pooling for convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2921–2930.
[40] H. Zheng, J. Fu, Z.-J. Zha, and J. Luo, "Learning deep bilinear transformation for fine-grained image representation," in *Advances in Neural Information Processing Systems*, 2019, pp. 4277–4286.
[41] E. Ustinova, Y. Ganin, and V. Lempitsky, "Multi-region bilinear convolutional neural networks for person re-identification," in *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2017, pp. 1–6.
[42] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 371–381.
[43] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 317–326.
[44] J.-H. Kim, K. W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," in *The 5th International Conference on Learning Representations*, 2017.
[45] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2008, pp. 1177–1184.
[46] P. Kar and H. Karnick, "Random feature maps for dot product kernels," in *Artificial Intelligence and Statistics*, 2012, pp. 583–591.
[47] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 239–247.
[48] H. Avron, H. Nguyen, and D. Woodruff, "Subspace embeddings for the polynomial kernel," in *Advances in neural information processing systems*, 2014, pp. 2258–2266.
[49] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.
[50] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2285–2294.
[51] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 365–381.
[52] B. N. Xia, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3760–3769.
[53] G. Chen, C. Lin, L. Ren, J. Lu, and J. Zhou, "Self-critical attention learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9637–9646.
[54] B. Scholkopf and A. J. Smola, "Learning with kernels: support vector machines, regularization, optimization, and beyond." Adaptive Computation and Machine Learning series, 2018.
[55] M. Lin, Q. Chen, and S. Yan, "Network in network," in *International Conference on Learning Representations*, 2013.
[56] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," vol. 3, no. 3. Distill, 2018, p. e10.
[57] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," in *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5. IEEE, 2010, pp. 978–994.
[58] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[60] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3960–3969.
[61] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9317–9326.
[62] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3702–3712.
[63] X. Jin, C. Lan, W. Zeng, G. Wei, and Z. Chen, "Semantics-aligned representation learning for person re-identification." in *AAAI*, 2020, pp. 11 173–11 180.
[64] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3186–3195.
[65] J. Liu, Y. Sun, C. Han, Z. Dou, and W. Li, "Deep representation learning on long-tailed data: A learnable embedding augmentation perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2970–2979.
[66] L. He and W. Liu, "Guided saliency feature learning for person re-identification in crowded scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 357–373.
[67] S. Zhao, C. Gao, J. Zhang, H. Cheng, C. Han, X. Jiang, X. Guo, W.-S. Zheng, N. Sang, and X. Sun, "Do not disturb me: Person re-identification under the interference of other pedestrians," in *European Conference on Computer Vision*. Springer, 2020, pp. 647–663.
[68] K. Zhu, H. Guo, Z. Liu, M. Tang, and J. Wang, "Identity-guided human semantic parsing for person re-identification," in *European Conference on Computer Vision*. Springer, 2020, pp. 346–363.
[69] Q. Zhou, B. Zhong, X. Liu, and R. Ji, "Attention-based neural architecture search for person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
[70] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Feature completion for occluded person re-identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4894–4912, 2022.
[71] 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.
[72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
[73] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.
[74] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," in *arXiv preprint arXiv:1703.07737*, 2017.

[75] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[76] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[77] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1821–1830.

[78] Z. Yu, J. Yu, C. Xiang, J. Fan, and D. Tao, "Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering," in *IEEE transactions on neural networks and learning systems*, vol. 29, no. 12. IEEE, 2018, pp. 5947–5959.