HOReID: Deep High-Order Mapping Enhances Pose Alignment for Person Re-Identification

Pingyu Wang, Zhicheng Zhao, Fei Su, Xingyu Zu, Nikolaos V. Boulgouris

***Abstract*—Despite the remarkable progress in recent years, person Re-Identification (ReID) approaches frequently fail in cases where the semantic body parts are misaligned between the detected human boxes. To mitigate such cases, we propose a novel *High-Order ReID* (HOReID) framework that enables semantic pose alignment by aggregating the fine-grained part details of multilevel feature maps. The HOReID adopts a high-order mapping of multilevel feature similarities in order to emphasize the differences of the similarities between aligned and misaligned part pairs in two person images. Since the similarities of misaligned part pairs are reduced, the HOReID enhances pose-robustness within the learned features. We show that our method derives from an intuitive and interpretable motivation and elegantly reduces the misalignment problem without using any prior knowledge from human pose annotations or pose estimation networks. This paper theoretically and experimentally demonstrates the effectiveness of the proposed HOReID, achieving superior performance over the state-of-the-art methods on the four large-scale person ReID datasets.**

***Index Terms*—Person Re-Identification, Pose Alignment, High-Order Mapping, Convolutional Neural Networks**

I. Introduction

PERSON Re-Identification (ReID) is the challenging task of matching person images of the same person across multiple cameras. Although the recent application of *Convolutional Neural Networks* (CNNs) on person ReID has been a great success, the problem of pose misalignment is far from being resolved, as shown in Fig. 1. In order to address this issue, prior ReID works have broadly followed two main paradigms, *i.e.*, pose-based [1–11] and pose-free [12–19] methods. The pose-based approaches require human pose annotations [2, 8, 10] or pose estimation networks [20–22] to supervise pose alignment. However, such high dependence on pose knowledge might limit the generalization of the ReID models to new person images with unseen human poses. Another widely-used workaround is to partition the person image into a few fixed rigid parts and learn detailed local features to get rid of human poses. Nevertheless, such coarse partition is unable to effectively align body parts without considering fine-grained pose variations within each part.

Pingyu Wang, Zhicheng Zhao, Fei Su and Xingyu Zu 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; zhaozc@bupt.edu.cn; sufei@bupt.edu.cn; zuxingyu@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) and MoE-CMCC Artifical Intelligence Project No. MCM20190701. (The first two authors contributed equally to this paper.)

In this paper, we propose a *High-Order ReID* (HOReID) framework to perform refined pose alignment by adaptively selecting aligned part pairs for computing image similarities. In the HOReID, we adopt a hierarchical structure to design two novel feature pooling layers, *i.e.*, a *Global Hierarchical Pooling* (GHP) layer and a *Local Hierarchical Pooling* (LHP) layer. Both the GHP and LHP layers learn high-order features to facilitate fine-grained pose alignment without relying on any additional pose estimation networks or human pose landmarks. Specifically, the GHP layer aims to align global image-based features including foreground and background regions. For excluding background effects, the LHP layer adopts a hierarchy-shared sampler to automatically sample discriminative part descriptors and then align the local body-based parts to learn subtler pose-robust features. Besides, we put forward a novel regularizer called *Discriminative Sampler Regularization* (DSR) to enrich the diversity of sampled body parts and dislodge background-aware regions. Since both the GHP and LHP layers compute the high-order mapping of multilevel feature maps, the image matching similarity is equivalent to an averaging sum of products of multilevel part similarities from all part pairs. According to the increasing property of the high-order mapping, the similarity product of aligned part pairs is dramatically larger than the misaligned part pairs. In this way, the HOReID highlights the aggregated similarities of the aligned part pairs without depending on the relative positions of body parts in person images. Therefore the pose misalignment problem is alleviated.

Overall, this paper makes the following contributions:

- We prove that the high-order mapping of multilevel feature similarities facilitates fine-grained pose alignment in theory.
- We propose an end-to-end HOReID framework with two novel layers, *i.e.*, GHP and LHP, which can learn both global and local pose-invariant representations.
- We propose a novel DSR regularizer to guide a descriptor sampler to detect discriminative body parts without using the supervision of pose landmark annotations.
- Our framework conducts pose alignment at the position level without relying on any extra prior pose knowledge, making it computationally efficient and highly generalizable to other unknown pose variations.
- The HOReID achieves state-of-the-art ReID performance on Market1501 [23], CUHK03 [24], DukeMTMC [25] and MSMT17 [26] datasets.

The remainder of this paper is organized as follows: In Sec. II, some related works about person ReID and bilinear

Fig. 1: Pose misalignment in person ReID caused by different camera viewpoints, different poses, imperfect person detection and partial body occlusion.

pooling are discussed. In Sec. III and IV, we introduce the details of the motivations and the proposed methods. In Sec. V and VI, the experimental discussions and results are reported. In Sec. VII, conclusions are drawn.

II. Related Work

A. Pose-Based Person Re-Identification

To solve the pose misalignment problem, human semantics in terms of pose/part is widely used to localize body parts for pose-robust feature learning [1–11]. For instance, GAN-based works [8, 10] use extra pose annotations to guide the generative model [27] to synthesize pose-specific person images and supervise the identity encoder model to mine pose-aligned features. In [28], body parts are first detected by DeeperCut [20] and multiple CNNs are designed for both global and local representation learning. Global and local features are aggregated by concatenation [2, 28] or mixture using a fully-connected layer [29]. Two stream networks, *i.e.*, GoogLeNet [30] and OpenPose [21], are applied in [9] to independently generate appearance and pose representations which are fused to intensify pose-robustness property within the learned descriptors. To achieve a more precise alignment, the fine-grained pixel-level person semantics predicted by DensePose [22] are used in [11] as an additional regularizer to guide the pose-robust representation learning from the original images. Some works [5, 7, 31] rely on constrained attention selection mechanisms from human pose information to implicitly align person representations by screening out semantic descriptors on the feature space. All of the above works aim to address misalignment using extra prior pose information. However it is non-trivial to obtain sufficient pose-labeled person images and strong pose estimation networks in real-world circumstances. Therefore, those methods might not generalize well to new images with unseen pose variations. In this work, the HOReID heads from a totally different but simple idea that highlights the similarities of aligned part pairs via the high-order mapping of multilevel feature similarities. Besides, our method is able to automatically rectify pose misalignment without depending on any extra pose knowledge, so it has increased practical significance and wide application prospect.

B. Pose-Free Person Re-Identification

Since pose annotations and networks are rarely available to the real-world applications, many works [12–19] adopt less-than-ideal alternative methods to align person features of body patches in a very coarse manner. As global features, which are learned from the full image, intend to capture the coarse-grained clues of appearance, the global feature maps in [13–15] are equally divided into multiple horizontal patches to exploit fine-grained local details. Analogously, the image feature map in [12] is rigidly partitioned into local stripes and a shortest path loss is introduced to align local stripes. Besides, some cross-modality person ReID works [32–34] also adopt the feature partition strategy to reduce the pose misalignment problem. The STN module [35] is integrated into ReID models [16–19] for rectifying person patches via affine transformation. However, the pose-free ReID approaches only achieve the coarse-grained pose alignment without considering detailed part semantics. In this work, we propose a subtle pose alignment method to achieve the efficient learning of semantically aligned features.

C. Bilinear Pooling

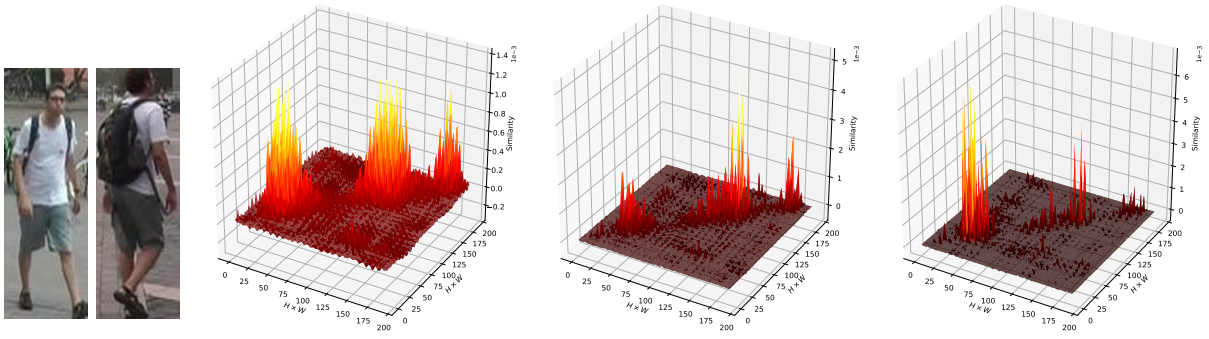
Bilinear pooling generates aggregate feature representations by using the Kronecker product. Such representations have achieved excellent performance in various visual recognition tasks [36–40]. For example, Lin *et al*. [36] propose bilinear pooling to aggregate the second-order feature interactions. For higher-order representation learning, Cui *et al*. [39] put forward a general pooling framework that captures higher-order interactions of features in the form of kernels. Furthermore, Cai *et al*. [40] propose a polynomial kernel based predictor to capture higher-order statistics of convolutional actions for modeling part interactions. For person Re-ID, Ustinova *et al*. [41] propose an architecture based on the deep bilinear convolutional network. Although that architecture leads to some performance improvements, it is not explicitly concerned with pose alignment. In this work, we extract multilevel appearance feature maps from a single backbone network. The extracted feature maps are then fused by the GHP and LHP layers in order to learn pose-aligned feature maps, without using any pose estimation networks. Besides, we extend the second-order bilinear pooling to the high-order pooling to explore the effectiveness of the high-order features on pose alignment.

III. Intuition and Motivation

A. Problem Definition

Given a training dataset of person images $\mathcal{D}_{\text{train}} = \{\boldsymbol{I}_i, \boldsymbol{y}_i\}_{i=1}^{N}$, where $\boldsymbol{I}_i$ and $\boldsymbol{y}_i$ are the ith person's image sample and identity (class) label respectively, and N is the number of training person images. During the training phase, we train a CNN model $\mathcal{F}(\cdot)$ to extract the convolutional feature $\boldsymbol{A} \in \mathbb{R}^{C\times H\times W}$ of the image $\boldsymbol{I}_a$, *i.e.*, $\boldsymbol{A} = \mathcal{F}(\boldsymbol{I}_a)$, where C, H and W denote the channel, height and width dimension, respectively. Then the feature map is pooled by a *Global Average Pooling* (GAP) layer to obtain the corresponding descriptor $\boldsymbol{a} \in \mathbb{R}^{C}$ as follows,

$$\boldsymbol{a} = \frac{1}{HW}\sum_{h_a w_a} \boldsymbol{A}^{h_a w_a}, \tag{1}$$



(a) $n = 1$, $\frac{\max}{\min} = \frac{1.42\times10^{-3}}{-3.30\times10^{-4}}$ **(b)** $n = 2$, $\frac{\max}{\min} = \frac{5.26\times10^{-3}}{-3.43\times10^{-4}}$ **(c)** $n = 3$, $\frac{\max}{\min} = \frac{7.09\times10^{-3}}{-3.51\times10^{-4}}$

Fig. 2: Similarity distribution of H^2W^2 part pairs of two images with the same class. We extract a pair of high-order feature maps from the first-order ($n = 1$), second-order ($n = 2$) and third-order ($n = 3$) GHP layers. The two feature maps are individually l_2 normalized by dividing the 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 high-order part similarities, respectively.

where $\boldsymbol{A}^{h_a w_a} \in \mathbb{R}^C$ is the part descriptor at the position (h_a, w_a). Besides, the part position satisfies $h_a \in [1, H]$ and $w_a \in [1, W]$. In the testing stage, given a pair of person images $\{\boldsymbol{I}_a, \boldsymbol{I}_b\}$ from a testing dataset $\mathcal{D}_{\text{test}}$, where the person classes are not overlapping between both datasets, we use CNNs in order to extract a pair of descriptor vectors $\{\boldsymbol{a}, \boldsymbol{b}\}$. Finally, the identity similarity between $\boldsymbol{a}$ and $\boldsymbol{b}$ is computed to judge whether $\boldsymbol{y}_a = \boldsymbol{y}_b$ or $\boldsymbol{y}_a \neq \boldsymbol{y}_b$.

B. Pose Misalignment

In this part, we give a mathematical treatment for the motivation of our HOReID framework. Suppose the two input person images $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$ from the same class, we use the inner product of $\boldsymbol{a}$ and $\boldsymbol{b}$ to measure the identity similarity of the two images $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$,

$$\begin{aligned}\text{Sim}\left(\boldsymbol{I}_a, \boldsymbol{I}_b\right) &= \left\langle \frac{1}{HW}\sum_{h_a w_a} \boldsymbol{A}^{h_a w_a}, \frac{1}{HW}\sum_{h_b w_b} \boldsymbol{B}^{h_b w_b} \right\rangle \\ &= \frac{1}{H^2W^2}\sum_{h_a w_a}\sum_{h_b w_b} \left\langle \boldsymbol{A}^{h_a w_a}, \boldsymbol{B}^{h_b w_b} \right\rangle\end{aligned}, \quad (2)$$

where $\langle \boldsymbol{a}, \boldsymbol{b} \rangle$ denotes the inner product between $\boldsymbol{a}$ and $\boldsymbol{b}$. In addition, $\boldsymbol{A}$ and $\boldsymbol{B}$ represent the convolutional feature maps of $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$, respectively. The similarity of $\boldsymbol{a}$ and $\boldsymbol{b}$ can be interpreted as an average sum of part similarities between H^2W^2 position pairs. However, such coarse similarity aggregation may degenerate into a suboptimal solution, which can be attributed to two major reasons. (1) The first reason is associated with the unbalanced quantity distribution between about HW aligned and $HW\,(HW-1)$ misaligned body part pairs. Since the number of the misaligned pairs (shoulder $\leftrightarrow$ hand) is quadratically larger than 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 person misalignment problem to some extent. (2) The second reason is related to the non-person part descriptors containing various background clutters as shown in Fig. 1. This problem is especially obvious when person bodies are partially occluded by other non-person objects. As a result, the background-aware part descriptors may bring an objectionable bias to the aggregated similarities in Eq. 2. In view of the two reasons, we propose two novel ideas, *i.e.*, highlighting aligned similarity and sampling semantic descriptor, to enhance the generalization capability of the ReID model.

C. Highlight Aligned Similarity

Motivated by the attention selection mechanism [42–44], we use the similarities from other hierarchical or non-hierarchical layers as attention values to intensify the aligned part similarities. Interestingly, we find that the hierarchical layers always outperform the non-hierarchical layers in our experiments of Sec. VI, because the hierarchical layers provide richer multi-level characteristics of person poses than the non-hierarchical ones. For adaptively weighting part similarities, we extend the first-order similarity to the high-order one by extracting n hierarchical feature maps $\{\boldsymbol{A}_l\}_{l=1}^n$ and $\{\boldsymbol{B}_l\}_{l=1}^n$,

$$\text{Sim}\left(\boldsymbol{I}_a, \boldsymbol{I}_b; n\right) = \frac{1}{H^2W^2}\sum_{h_a w_a}\sum_{h_b w_b}\prod_{l=1}^{n}\left\langle \boldsymbol{A}_l^{h_a w_a}, \boldsymbol{B}_l^{h_b w_b} \right\rangle, \quad (3)$$

where $\boldsymbol{A}_l, \boldsymbol{B}_l \in \mathbb{R}^{C\times H\times W}$ denote the l-th level feature maps. The high-order part similarity can be represented as the product of the similarities of all n part similarities in the same position pairs. Since the aligned body parts usually contain identical semantics, the similarity of the aligned part pair is likely to be higher than the similarity of the misaligned part pair at the same feature level. To make the part similarity be always non-negative, we restrict $\boldsymbol{A}_l$ and $\boldsymbol{B}_l$ to be element-wise non-negative by adding a ReLU layer. In this way, the high-order part similarity may lead to an exponential increase of the similarity discrepancies between aligned and misaligned part pairs in theory. Notably, the high-order product results in a sharper distribution of part feature similarities as shown in Fig. 2. Hence the high-order mapping highlights the importance of aligned part pairs without depending on the relative positions of person parts in person images. As the order n increases, the aggregated similarity is approximated by the summation of the similarity product from the aligned pair set. Therefore, the high-order similarity is beneficial to solve the pose misalignment problem without the requirement of auxiliary person pose knowledge during the training and testing phase.

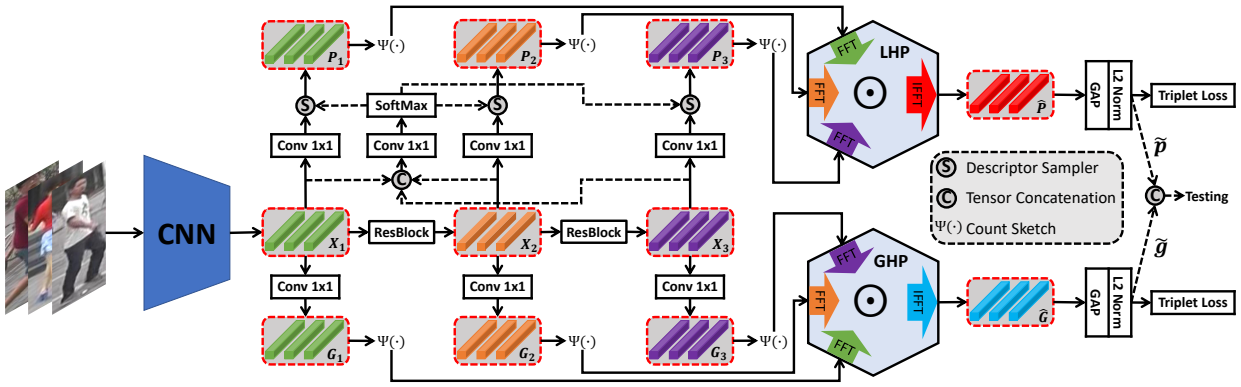


Fig. 3: Overview of the proposed HOReID framework. It consists of three parts, *i.e.*, a backbone network, a *Global Hierarchical Pooling* (GHP) layer and a *Local Hierarchical Pooling* (LHP) layer. The backbone network is input with misaligned person images to extract convolutional person representations. Then we adopt a series of nonlinear residual blocks to produce multilevel convolutional feature maps with informative pose knowledge. Next, those multilevel feature maps are fed into the GHP and LHP layer to output high-order global and local features. They are supervised by two independent triplet loss functions during the training stage, while we concatenate them to make use of the global and local information during the testing phase.

D. Sample Semantic Descriptor

Since the background descriptors are likely similar, the aligned part pair set may still contain several non-person part locations. Then aligned background similarity brings some noise to the similarity aggregation, which hampers background-robust feature learning. Thus, the foreground descriptors should be sampled into Eq. 3 and the background descriptors need to be excluded from global features. Meanwhile, each high-order part similarity is calculated using the product of multilevel similarities from the same part pairs. For this reason, the sampler should select the same part locations for all levels. To realize this goal, an identical sampler $\mathbb{S}$ needs to be shared among multilevel feature maps, otherwise the phenomenon of interlevel inconsistency occurs, implying that local descriptors are sampled inappropriately. The shared sampler $\mathbb{S}$ is formulated by,

$$\begin{aligned}\overline{\boldsymbol{A}}_l &= \left\{\mathbb{S}\left(h_a, w_a; \{\boldsymbol{A}_{l'}\}_{l'=1}^{n}\right)\boldsymbol{A}_l^{h_a w_a}\right\},\\ \overline{\boldsymbol{B}}_l &= \left\{\mathbb{S}\left(h_b, w_b; \{\boldsymbol{B}_{l'}\}_{l'=1}^{n}\right)\boldsymbol{B}_l^{h_b w_b}\right\},\end{aligned} \tag{4}$$

where $h_a, h_b \in [1, H]$ and $w_a, w_b \in [1, W]$. $\overline{\boldsymbol{A}}_l, \overline{\boldsymbol{B}}_l \in \mathbb{R}^{C\times P}$ denote the set of P sampled descriptors at the l-th level. $\mathbb{S}(h, w)$ returns 1 if the part descriptor at position (h, w) is sampled, otherwise 0. Hence, Eq. 3 is rewritten as

$$\mathrm{Sim}\left(\boldsymbol{I}_a, \boldsymbol{I}_b; n\right) = \frac{1}{P^2}\sum_{p_a}\sum_{p_b}\prod_{l=1}^{n}\left\langle \overline{\boldsymbol{A}}_l^{p_a}, \overline{\boldsymbol{B}}_l^{p_b}\right\rangle, \tag{5}$$

where $\overline{\boldsymbol{A}}_l^{p_a}$ and $\overline{\boldsymbol{B}}_l^{p_b}$ denote the p_a-th and p_b-th descriptors of $\overline{\boldsymbol{A}}_l$ and $\overline{\boldsymbol{B}}_l$.

Theorem 1. *Suppose* $\boldsymbol{x}^{(n)} = Vec\left(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2 \otimes \cdots \otimes \boldsymbol{x}_n\right)$ *and* $\boldsymbol{y}^{(n)} = Vec\left(\boldsymbol{y}_1 \otimes \boldsymbol{y}_2 \otimes \cdots \otimes \boldsymbol{y}_n\right)$ *are two high-order vectors generated by Kronecker product* $\otimes$ *with* n *input features* $\{\boldsymbol{x}_l\}_{l=1}^{n}$ *and* $\{\boldsymbol{y}_l\}_{l=1}^{n}$*, where* $Vec(\cdot)$ *transforms a tensor to vector. The similarity of two high-order vectors* $\boldsymbol{x}^{(n)}$ *and* $\boldsymbol{y}^{(n)}$ *is equivalent to the product of all similarities of* n *input vectors,* $\left\langle \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right\rangle = \prod_{l=1}^{n}\left\langle \boldsymbol{x}_l, \boldsymbol{y}_l\right\rangle$

Proof. Proof is provided with mathematical induction.

Base case: When $n = 2$, we have two vectors generated by the Kronecker product: $\boldsymbol{x}^{(2)} = \mathrm{Vec}\left(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2\right)$ and $\boldsymbol{y}^{(2)} = \mathrm{Vec}\left(\boldsymbol{y}_1 \otimes \boldsymbol{y}_2\right)$, where $\boldsymbol{x}_1, \boldsymbol{y}_1 \in \mathbb{R}^{D_1}$, $\boldsymbol{x}_2, \boldsymbol{y}_2 \in \mathbb{R}^{D_2}$ and $\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)} \in \mathbb{R}^{D_1 D_2}$. The similarity of $\boldsymbol{x}^{(2)}$ and $\boldsymbol{y}^{(2)}$ is computed by the innner product,

$$\begin{aligned}\left\langle \boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}\right\rangle &= \left\langle \mathrm{Vec}\left(\boldsymbol{x}_1 \otimes \boldsymbol{x}_2\right), \mathrm{Vec}\left(\boldsymbol{y}_1 \otimes \boldsymbol{y}_2\right)\right\rangle\\ &= \sum_{d_1=1}^{D_1}\sum_{d_2=1}^{D_2}\left(\boldsymbol{x}_1\left[d_1\right]\cdot\boldsymbol{x}_2\left[d_2\right]\right)\cdot\left(\boldsymbol{y}_1\left[d_1\right]\cdot\boldsymbol{y}_2\left[d_2\right]\right)\\ &= \left(\sum_{d_1=1}^{D_1}\boldsymbol{x}_1\left[d_1\right]\cdot\boldsymbol{y}_1\left[d_1\right]\right)\cdot\left(\sum_{d_2=1}^{D_2}\boldsymbol{x}_2\left[d_2\right]\cdot\boldsymbol{y}_2\left[d_2\right]\right)\\ &= \prod_{l=1}^{2}\left\langle \boldsymbol{x}_l, \boldsymbol{y}_l\right\rangle\end{aligned}, \tag{6}$$

where $\boldsymbol{x}_1\left[d_1\right]$ and $\boldsymbol{x}_2\left[d_2\right]$ denote the d_1-th and d_2-th entry of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Therefore the statement is correct when $n = 2$.

Induction step: When $n > 2$, we suppose $\left\langle \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right\rangle = \prod_{l=1}^{n}\left\langle \boldsymbol{x}_l, \boldsymbol{y}_l\right\rangle$. We need to prove $\left\langle \boldsymbol{x}^{(n+1)}, \boldsymbol{y}^{(n+1)}\right\rangle = \prod_{l=1}^{n+1}\left\langle \boldsymbol{x}_l, \boldsymbol{y}_l\right\rangle$. According to Eq. 6, we have:

$$\begin{aligned}\left\langle \boldsymbol{x}^{(n+1)}, \boldsymbol{y}^{(n+1)}\right\rangle &= \left\langle \mathrm{Vec}\left(\boldsymbol{x}^{(n)}\otimes\boldsymbol{x}_{n+1}\right), \mathrm{Vec}\left(\boldsymbol{y}^{(n)}\otimes\boldsymbol{y}_{n+1}\right)\right\rangle\\ &= \left\langle \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right\rangle\left\langle \boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}\right\rangle\\ &= \prod_{l=1}^{n}\left\langle \boldsymbol{x}_l, \boldsymbol{y}_l\right\rangle\cdot\left\langle \boldsymbol{x}_{n+1}, \boldsymbol{y}_{n+1}\right\rangle = \prod_{l=1}^{n+1}\left\langle \boldsymbol{x}_l, \boldsymbol{y}_l\right\rangle\end{aligned}. \tag{7}$$

As seen, the statement is correct when $n > 2$. Hence by mathematical induction $\left\langle \boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}\right\rangle = \prod_{l=1}^{n}\left\langle \boldsymbol{x}_l, \boldsymbol{y}_l\right\rangle$ is correct for $n \geq 2$ and the proof is completed. □

IV. Proposed Method

In this section, we first overview the hierarchical structure of the HOReID framework shown in Fig. 3, then we introduce the *Global Hierarchical Pooling* (GHP) layer and *Local Hierarchical Pooling* (LHP) layer, which are key components of the HOReID framework.

A. Hierarchical Network Architecture

Given an input person image $\boldsymbol{I}_x$, we extract the first-level hidden feature map $\boldsymbol{X}_1 \in \mathbb{R}^{C\times H\times W}$ from the backbone network. Then $n-1$ cascaded residual blocks [46], containing an 1×1 convolutional layer, a BatchNorm [47] layer and a ReLU layer, are used to generate n-level hidden feature maps $\{\boldsymbol{X}_l\}_{l=1}^{n}$ of the same size. To encode multilevel global information,

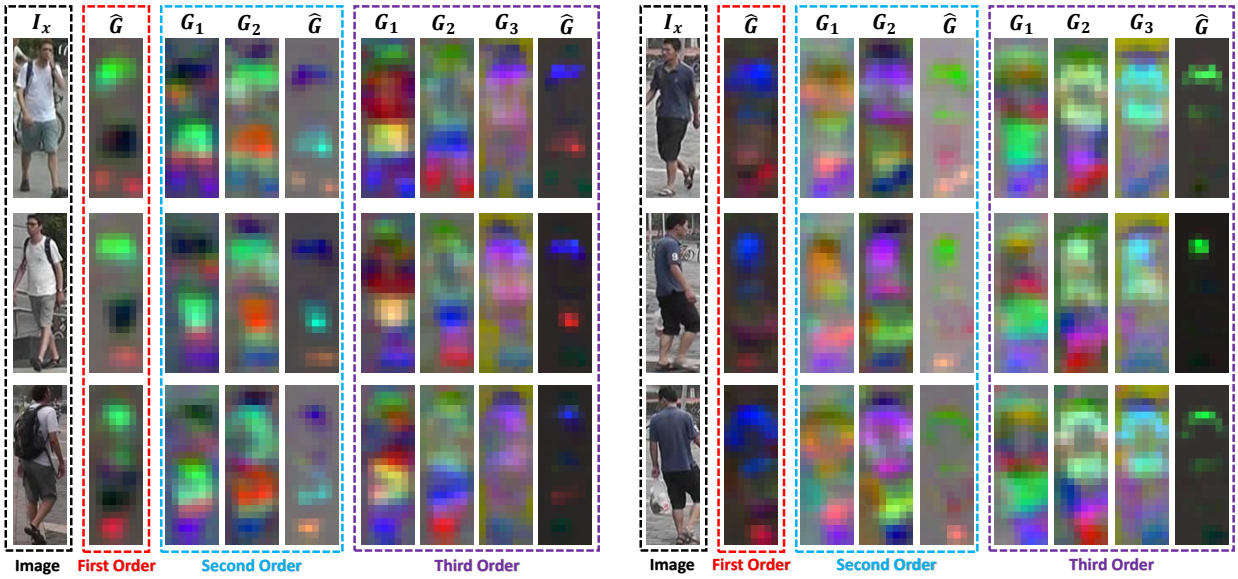


Fig. 4: Visualization of multilevel feature maps $\{\boldsymbol{G}_l\}_{l=1}^n$ and high-order feature maps $\widehat{\boldsymbol{G}}$ extracted from the first-order ($n=1$), second-order ($n=2$) and third-order ($n=3$) GHP layer. Following the SIFTFlow [45], we use *Principal Component Analysis* (PCA) to transform all part features to 3D vectors and then normalize 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 different.

we use an 1×1 convolutional layer which obtains image-based global feature maps $\boldsymbol{G}_l \in \mathbb{R}^{C_g \times H \times W}$ from $\boldsymbol{X}_l$, where C_g is the channel dimension. Besides, we generate body-based part descriptors $\boldsymbol{P}_l \in \mathbb{R}^{C_p \times P}$ from $\boldsymbol{X}_l$ to eliminate the background descriptors via an effective sampling algorithm $\mathbb{S}$, where C_p is the channel dimension. All global and local feature maps, *i.e.*, $\{\boldsymbol{G}_l\}_{l=1}^n$ and $\{\boldsymbol{P}_l\}_{l=1}^n$, are aggregated by the GHP and LHP layers in order to capture high-order global and local alignment interactions, respectively. Finally, both global and local representations are concatenated along the feature dimension for evaluation.

B. Global Hierarchical Pooling

Global Feature Aggregation: According to the analysis in Sec. III, the high-order similarity aggregation contributes to reducing the pose misalignment problem. However, calculating high-order similarities with Eq. 3 has large computational cost and memory consumption during the inference. For example, the two convolutional features, *i.e.*, $\boldsymbol{A}$ and $\boldsymbol{B}$, are required to be stored in memory, which means that the space cost is at least $\mathcal{O}(C_g HW)$. Furthermore, we need to compute the high-order similarities of H^2W^2 part pairs, for which the associated time cost is at least $\mathcal{O}\left(C_g H^2 W^2\right)$. In order to mitigate this issue, we need to aggregate these convolutional feature maps into a single feature vector. According to Theorem 1, the multilevel features can be aggregated into a high-order descriptor vector by the Kronecker product $\otimes$, while the similarities of the high-order descriptors are equivalent to the high-order similarity summation of H^2W^2 part pairs in Eq. 3. For the n-level part descriptors $\left\{\boldsymbol{G}_l^{hw}\right\}_{l=1}^n$, the high-order feature is generated by

$$\boldsymbol{G}^{hw} = \mathrm{Vec}\left(\boldsymbol{G}_1^{hw} \otimes \boldsymbol{G}_2^{hw} \otimes \cdots \otimes \boldsymbol{G}_n^{hw}\right), \tag{8}$$

where $\mathrm{Vec}(\cdot)$ transforms a tensor to a vector. The Kronecker product allows all elements of n feature vectors to interact with each other and therefore exhibits strong representational capabilities. However, the channel dimension of the high-order feature maps increases exponentially, leading to very high memory consumption $\mathcal{O}\left(HWC_g^n\right)$ and computational complexity $\mathcal{O}\left(HWC_g^n\right)$.

Compact Embedding Approximation: We thus need an effective and efficient approach that projects the Kronecker product to a lower dimensional space and also evades computing the Kronecker product directly. As suggested by Pagh [48], a differentiable function named count sketch $\boldsymbol{q} = \Psi(\boldsymbol{p})$ can be applied to project the high-dimensional vector $\boldsymbol{p}$ to the low-dimensional vector $\boldsymbol{q}$ without significantly degrading the representational capability. To avoid the Kronecker product explicitly, the previous work [37, 48, 49] has demonstrated that the count sketch of the Kronecker product with n vectors is equivalent to the convolution of their count sketches,

$$\Psi\left(\boldsymbol{G}^{hw}\right) = \Psi\left(\boldsymbol{G}_1^{hw}\right) \circledast \Psi\left(\boldsymbol{G}_2^{hw}\right) \circledast \cdots \circledast \Psi\left(\boldsymbol{G}_n^{hw}\right), \tag{9}$$

where $\circledast$ is a convolution operation. According to the convolution theorem, convolution in the space/time domain is equivalent to Hadamard product $\odot$ in the frequency domain,

$$\begin{aligned}\widehat{\boldsymbol{G}}^{hw} &= \Psi\left(\boldsymbol{G}^{hw}\right) \\ &= \mathrm{IFFT}\left(\mathrm{FFT}\left(\Psi\left(\boldsymbol{G}_1^{hw}\right)\right) \odot \cdots \odot \mathrm{FFT}\left(\Psi\left(\boldsymbol{G}_n^{hw}\right)\right)\right),\end{aligned} \tag{10}$$

where $\mathrm{FFT}(\cdot)$ and $\mathrm{IFFT}(\cdot)$ denote original and inverse fast Fourier transformation, respectively. Besides, Eq. 10 effectively and efficiently aggregates multilevel features because the combination takes place in the form of Hadamard product. In this way, we obtain a low-dimensional representation $\widehat{\boldsymbol{G}}^{hw} \in \mathbb{R}^{D_g}$ to approximate $\boldsymbol{G}^{hw} \in \mathbb{R}^{C_g^n}$, where the compressed dimension D_g satisfies $D_g \ll C_g^n$. After applying the proposed GHP layer to $H \times W$ spatial locations, we obtain an approximatively high-order feature map $\widehat{\boldsymbol{G}} \in \mathbb{R}^{D_g \times H \times W}$ with the low time cost $\mathcal{O}\left(nHWC_g + nHWD_g \log D_g\right)$ and space complexity $\mathcal{O}\left(nHWC_g + nC_gD_g\right)$. Subsequently, the

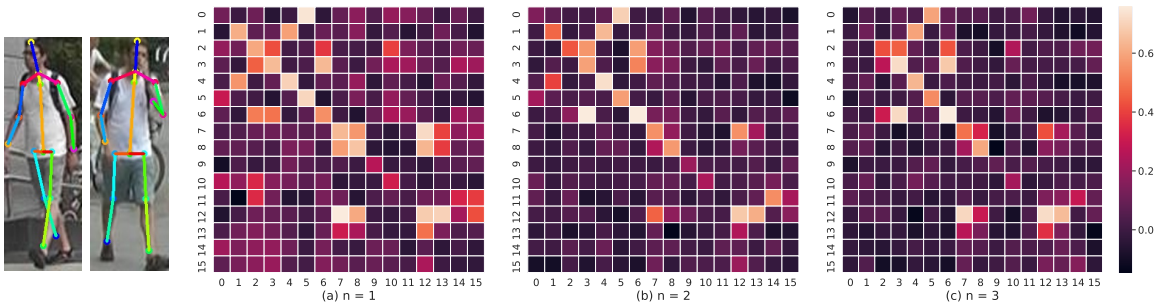


Fig. 5: Part similarity distribution of two images from the same classes. We use an existing OpenPose [21] to detect 16 body landmarks and then 16 part features are extracted from the first-order ($n = 1$), second-order ($n = 2$) and third-order ($n = 3$) GHP layers. For each GHP layer, we compute cosine similarities among 16×16 landmark pairs.

global feature vector $\widehat{\boldsymbol{g}} \in \mathbb{R}^{D_g}$ is calculated further by a GAP layer. Since the Hadamard product may make the feature norm aggrandize dramatically, the model might converge to a suboptimal local minimum. Therefore, a l_2 normalization ($\widetilde{\boldsymbol{g}} = \widehat{\boldsymbol{g}} / \left\|\widehat{\boldsymbol{g}}\right\|_2$) layer is appended after the pooled vector $\widehat{\boldsymbol{g}}$.

C. Local Hierarchical Pooling

Attentional Descriptor Sampler: As mentioned in Sec. III, the background descriptors should be excluded for background-invariant feature learning. Unfortunately, it is non-trivial to directly adopt the sampling method because the sampler $\mathbb{S}$ in Eq. 4 is untrainable with backpropagation algorithm. Motivated by attention learning [42–44], we propose to conduct an adaptive sampling approach via the softmax function. Suppose an 1×1 convolutional filter can be regarded as a semantic part detector, we can use P convolutional filters to generate P confidence maps,

$$\boldsymbol{M} = \phi\left(\{\boldsymbol{X}_l\}_{l=1}^{n}; \boldsymbol{W}_\phi\right), \tag{11}$$

where $\boldsymbol{M} \in \mathbb{R}^{P \times H \times W}$ and $\boldsymbol{W}_\phi$ denotes the parameter of convolutional filters. As the confidence map $\boldsymbol{M}^p \in \mathbb{R}^{H \times W}$ has high response to the p-th body part, the body part feature can be generated by the weighted average of all part features according to the confidence map. Additionally, the confidence map is normalized by the softmax function,

$$\boldsymbol{P}_l^p = \sum_{hw} \overline{\boldsymbol{G}}_l^{hw} \frac{\exp\left(\boldsymbol{M}^{phw}\right)}{\sum_{h'w'} \exp\left(\boldsymbol{M}^{ph'w'}\right)}, \tag{12}$$

where $\boldsymbol{M}^{phw}$ is the entry of the tensor $\boldsymbol{M}$ at the position (p, h, w) and $\boldsymbol{P}_l^p \in \mathbb{R}^{C_p}$ is the p-th sampled part descriptor of the global feature map $\overline{\boldsymbol{G}}_l \in \mathbb{R}^{C_p \times H \times W}$. Note that we do not share the parameters of the 1×1 convolutional layers for $\boldsymbol{G}_l$ and $\overline{\boldsymbol{G}}_l$. Compared with the sampling method in Eq. 4, this meticulous design of the shared sampler offers an effective way to gather key features: it captures the global person features (clothing and color), when $\boldsymbol{M}^p$ is densely attended on a large region; and it captures the local person parts (head and leg), when $\boldsymbol{M}^p$ is sparsely attended on a small area.

Discriminative Sampler Regularization: For sampling different semantic descriptors, the confidence map $\boldsymbol{M}^p$ should possess a discriminative probability distribution for each sampled descriptor. Besides, without any specific treatments, there is no guarantee that sampled descriptors contain different semantic information. In other words, multiple confidence maps could easily learn to detect the same body part. In practice, we need to ensure each confidence map focuses on different regions of the given image. One straightforward way to achieve this is to reduce the spatial overlap between different confidence maps. However, using discriminative confidence maps is still unable to precisely locate different body parts because it is very likely to sample similar descriptors from person parts with large regions, *e.g.*, clothing and trousers. To solve this problem, we put forward a simple yet effective regularizer named *Discriminative Sampler Regularization* (DSR) to reduce the similarities between different sampled descriptors. For a given pair of sampled descriptors $\boldsymbol{P}_l^i$ and $\boldsymbol{P}_l^j$ from the i-th and j-th body parts, we expect their corresponding cosine similarities to be smaller than a margin m_p,

$$\mathcal{L}_{\text{reg}} = \frac{1}{2nP(P-1)} \sum_{l=1}^{n} \sum_{i \neq j} \left[\left\langle \widetilde{\boldsymbol{P}}_l^i, \widetilde{\boldsymbol{P}}_l^j \right\rangle - m_p\right]_+, \tag{13}$$

where $\widetilde{\boldsymbol{P}}_l^i$ and $\widetilde{\boldsymbol{P}}_l^j$ denote l_2 normalized $\boldsymbol{P}_l^i$ and $\boldsymbol{P}_l^j$, respectively. $[\cdot]_+ = \max(\cdot, 0)$ denotes a hinge function and the margin m_p is set as $m_p = 0.2$.

Local Feature Aggregation: Following the GHP layer, the high-order part features are approximatively computed by

$$\widehat{\boldsymbol{P}}^p = \text{IFFT}\left(\text{FFT}\left(\Psi\left(\boldsymbol{P}_1^p\right)\right) \odot \cdots \odot \text{FFT}\left(\Psi\left(\boldsymbol{P}_n^p\right)\right)\right), \tag{14}$$

where $\widehat{\boldsymbol{P}}^p \in \mathbb{R}^{D_p}$ is a low-dimensional descriptor vector of the p-th sampled body part. Then all P part features $\widehat{\boldsymbol{P}} = \left[\widehat{\boldsymbol{P}}^1, \widehat{\boldsymbol{P}}^2, \cdots, \widehat{\boldsymbol{P}}^P\right] \in \mathbb{R}^{D_p \times P}$ are pooled by a GAP layer to obtain a feature vector as follows,

$$\widehat{\boldsymbol{p}} = \frac{1}{P} \sum_{p} \widehat{\boldsymbol{P}}^p. \tag{15}$$

Finally, a l_2 normalization ($\widetilde{\boldsymbol{p}} = \widehat{\boldsymbol{p}} / \left\|\widehat{\boldsymbol{p}}\right\|_2$) layer is utilized to reduce the magnitude variations caused by the Hadamard product.

D. Overall Loss Function

In order to train the HOReID framework, we utilize the triplet loss function. We define $\widetilde{\boldsymbol{g}}_a$, $\widetilde{\boldsymbol{g}}_p$ and $\widetilde{\boldsymbol{g}}_n$ as the anchor, positive and negative high-order global features, while $\widetilde{\boldsymbol{p}}_a$, $\widetilde{\boldsymbol{p}}_p$ and

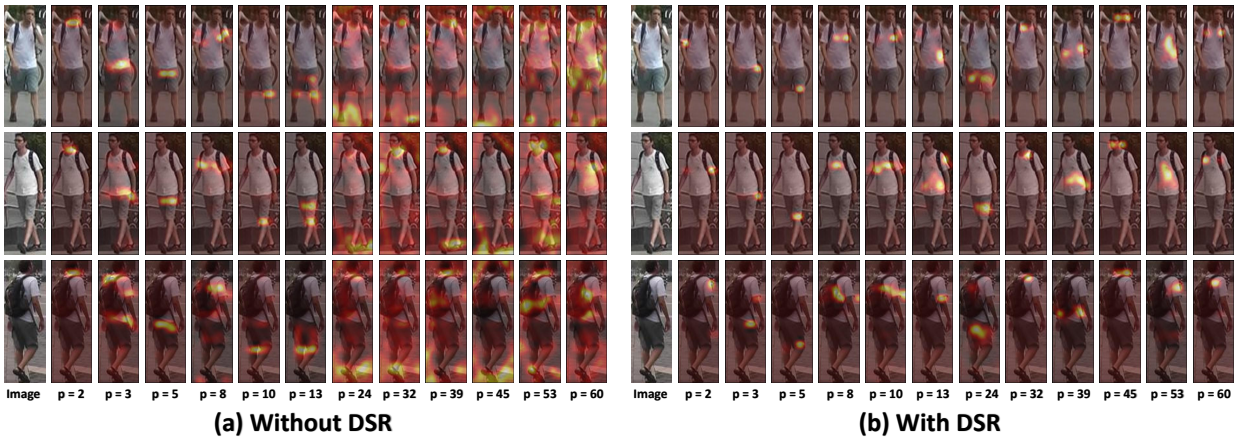


Fig. 6: Visualization of different confidence maps in the LHP layer. We extract a series of confidence maps from $\{2, 3, 5, 8, 10, 13, 24, 32, 39, 45, 53, 60\}$ channel dimensions, then their values are normalized by a softmax function. The body parts with large attention values are highlighted by bright colors.

$\widetilde{\boldsymbol{p}}_n$ represent the anchor, positive and negative high-order part features. The triplet loss function is formulated as:

$$\mathcal{L}_{\text{tri}} = \frac{1}{|\mathcal{S}|} \sum_{(a,p,n)\in\mathcal{S}} \Big[\langle \widetilde{\boldsymbol{g}}_a, \widetilde{\boldsymbol{g}}_n\rangle - \langle \widetilde{\boldsymbol{g}}_a, \widetilde{\boldsymbol{g}}_p\rangle + m_t\Big]_+ + \Big[\langle \widetilde{\boldsymbol{p}}_a, \widetilde{\boldsymbol{p}}_n\rangle - \langle \widetilde{\boldsymbol{p}}_a, \widetilde{\boldsymbol{p}}_p\rangle + m_t\Big]_+, \quad (16)$$

where the margin m_t is set as $m_t = 0.2$ and $\mathcal{S}$ is the set of $|\mathcal{S}|$ triplets within one mini-batch. The overall loss function is defined by $\mathcal{L} = \mathcal{L}_{\text{tri}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}$, where $\lambda_{\text{reg}} = 0.1$ is the loss weight of $\mathcal{L}_{\text{reg}}$.

V. Discussion

A. Feature Visualization

In the following, considering the collaborative effect of hierarchical interactions among multilevel features in Eq. 10, we give a microscopic interpretation from the perspective of feature visualization, which is also a strong justification of our method. In part, it reveals the reason why the high-order features are more pose-robust and discriminative than low-order features.

For a given input image $\boldsymbol{I}_x$, we extract multilevel feature maps $\{\boldsymbol{G}_l\}_{l=1}^n$ and high-order feature maps $\widehat{\boldsymbol{G}}$ from the first-order, second-order and third-order GHP layers. As shown in Fig. 4, one can observe that multi-level feature maps mainly encode the semantics of various body parts, including heads, hands and legs, and the corresponding colors differ depending on their spatial locations. Besides, the descriptors with the same location of multilevel feature maps exhibit different colors because of the diversity of hierarchical feature maps. In view of the color distribution, the high-order features concentrate on encoding the discriminative semantic parts (*e.g.*, heads and shoulders) to represent person identities, while the low-order features prefer to capture coarse global appearance information. Accordingly, the high-order information contributes to intensifying the pose-robustness and discrimination property within the learned person features.

B. Similarity Distribution

Based on the high-order similarity aggregation in Eq. 3, we give another macroscopic interpretation from the distribution of high-order part similarities. In a sense, it exhibits a valuable angle to illustrate the relationship between high-order feature similarity and fine-grained pose alignment.

Given two input images $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$ of the same classes, we extract a pair of aggregated convolutional feature maps $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{B}}$ from the first-order, second-order and third-order GHP layers. The two feature maps are individually l_2-normalized by dividing the norms of spatially pooled features $\widehat{\boldsymbol{a}}$ and $\widehat{\boldsymbol{b}}$. Finally, the similarity matrix is calculated by the inner product of all H^2W^2 part pairs, as shown in Fig. 2. We observe an interesting phenomenon that the maximum part similarity of the high-order features is significantly larger than the low-order features, while the minimum part similarity almost keeps unchanged for all orders. In addition, the amount of misaligned pairs with prominent similarities consistently decreases along with the increase of feature orders. In Fig. 5, the visualized results well demonstrate that the high-order part features can successfully learn landmark correspondences between two images without using pose annotations. Therefore, the high-order GHP layer alleviates the pose misalignment problem by increasing the similarities of aligned part pairs and decreasing the similarities of misaligned part pairs.

C. Sampler Visualization

In this part, we provide an intuitionistic interpretation by visualizing confidence maps to study the impact of the proposed descriptor sampler. The visualized results convincingly demonstrate the superiority of the proposed DSR. To some degree, the interpretation also clarifies the reason why learning body-based features is more beneficial to the pose alignment than learning image-based features.

Given an input person image $\boldsymbol{I}_x$, we use the LHP layer in order to extract a series of normalized confidence maps $\boldsymbol{M}^p$ ($p \in \{2, 3, 5, 8, 10, 13, 24, 32, 39, 45, 53, 60\}$). To better visualize the spatial relationships between confidence maps and body parts, the low-resolution map $\boldsymbol{M}^p$ is upsampled by the cubic interpolation to have the same size as $\boldsymbol{I}_x$ and then we merge both $\boldsymbol{M}^p$ and $\boldsymbol{I}_x$ by alpha blending, as shown in Fig. 6. In each visualized example, the bright areas represent sampled regions while the dark ones are not sampled in the LHP layer. On the whole, most of confidence maps focus on foreground regions and generally correspond to specific body parts. In Fig. 6 (a), without the DSR, some confidence maps $\boldsymbol{M}^p$

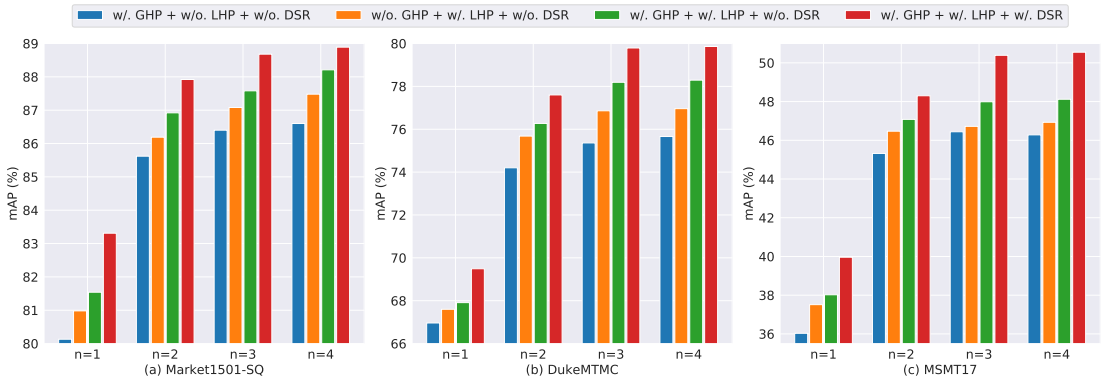


Fig. 7: Ablation study of the GHP and LHP layers with ResNet-50 on the Maket1501, DukeMTMC and MSMT17 datasets.

($p \in \{24, 32, 39, 45, 53, 60\}$) are unable to locate semantic body parts and prefer to mine background-aware information, which might hamper pose-robust feature learning. In Fig. 6 (b), with the DSR, all confidence maps not only significantly exclude background regions but also detect different body parts to enhance the diversity of sampled descriptors. To understand the contribution of the DSR in depth, we consider an extreme situation where two non-person part descriptors are accidentally sampled. Since the non-person descriptors usually contain background information, they can appear very similar in the feature space as shown in Fig. 4, which may violate the optimization objective of the DSR. Hence, to avoid the production of a large loss value, the DSR will tend to avoid situations where non-person part descriptors are sampled more than one time. Compared with pose-based methods [8–11], our methodology specializes in discovering semantically distinctive body parts without using any extra estimation networks or human pose annotations, which is useful for fine-grained pose alignment in person ReID.

D. Backpropagation Optimization

Considering the collaborative effect of high-order similarities, we can provide another theoretical analysis from the back-propagation optimization for the triplet loss.

For the simplification of the following analysis, we ignore the l_2 normalization for high-order descriptors. Suppose the high-order features are directly aggregated by the the Kronecker product without using count sketch approximation, so the triplet loss for the GHP layer is formulated as,

$$\begin{aligned}\mathcal{L} &= \Big[\langle \widehat{\boldsymbol{g}}_a, \widehat{\boldsymbol{g}}_n\rangle - \langle \widehat{\boldsymbol{g}}_a, \widehat{\boldsymbol{g}}_p\rangle + m_t\Big]_+ \\ &= \Big[\frac{1}{H^2W^2}\sum_{h_aw_a}\sum_{h_nw_n}\prod_{l=1}^{n_o}\langle \boldsymbol{G}_{al}^{h_aw_a}, \boldsymbol{G}_{nl}^{h_nw_n}\rangle \\ &\quad - \frac{1}{H^2W^2}\sum_{h_aw_a}\sum_{h_pw_p}\prod_{l=1}^{n_o}\langle \boldsymbol{G}_{al}^{h_aw_a}, \boldsymbol{G}_{pl}^{h_pw_p}\rangle + m_t\Big]_+\end{aligned}, \quad (17)$$

where $\boldsymbol{G}_{al}^{h_aw_a}$ denotes the part descriptor vector of the anchor feature map $\boldsymbol{G}_{al}$ at the position (h_a, w_a) of the l-th level. The similar definition is also adopted to the positive and negative feature maps, *i.e.*, $\boldsymbol{G}_{pl}^{h_pw_p}$ and $\boldsymbol{G}_{nl}^{h_nw_n}$. To optimize Eq. 17, we can calculate its gradient with respect to $\boldsymbol{G}_{al}^{h_aw_a}$, $\boldsymbol{G}_{pl}^{h_pw_p}$ and $\boldsymbol{G}_{nl}^{h_nw_n}$ respectively as,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial \boldsymbol{G}_{al}^{h_aw_a}} &= \frac{1}{H^2W^2}\sum_{h_nw_n}\boldsymbol{G}_{nl}^{h_nw_n}\prod_{l'\neq l}\langle \boldsymbol{G}_{al'}^{h_aw_a}, \boldsymbol{G}_{nl'}^{h_nw_n}\rangle \\ &\quad - \frac{1}{H^2W^2}\sum_{h_pw_p}\boldsymbol{G}_{pl}^{h_pw_p}\prod_{l'\neq l}\langle \boldsymbol{G}_{al'}^{h_aw_a}, \boldsymbol{G}_{pl'}^{h_pw_p}\rangle, \\ \frac{\partial\mathcal{L}}{\partial \boldsymbol{G}_{pl}^{h_pw_p}} &= -\frac{1}{H^2W^2}\sum_{h_aw_a}\boldsymbol{G}_{al}^{h_aw_a}\prod_{l'\neq l}\langle \boldsymbol{G}_{al'}^{h_aw_a}, \boldsymbol{G}_{pl'}^{h_pw_p}\rangle, \\ \frac{\partial\mathcal{L}}{\partial \boldsymbol{G}_{nl}^{h_nw_n}} &= \frac{1}{H^2W^2}\sum_{h_aw_a}\boldsymbol{G}_{al}^{h_aw_a}\prod_{l'\neq l}\langle \boldsymbol{G}_{al'}^{h_aw_a}, \boldsymbol{G}_{nl'}^{h_nw_n}\rangle,\end{aligned} \quad (18)$$

if the constraint of Eq. 17 is violated, or zero otherwise. Here, we define the following product formula by,

$$\begin{aligned}\boldsymbol{S}_{anl}^{h_aw_a,h_nw_n} &= \prod_{l'\neq l}\langle \boldsymbol{G}_{al'}^{h_aw_a}, \boldsymbol{G}_{nl'}^{h_nw_n}\rangle, \\ \boldsymbol{S}_{apl}^{h_aw_a,h_pw_p} &= \prod_{l'\neq l}\langle \boldsymbol{G}_{al'}^{h_aw_a}, \boldsymbol{G}_{pl'}^{h_pw_p}\rangle.\end{aligned} \quad (19)$$

Therefore, Eq. 18 is rewritten as,

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial \boldsymbol{G}_{al}^{h_aw_a}} &= \frac{1}{H^2W^2}\sum_{h_nw_n}\boldsymbol{S}_{anl}^{h_aw_a,h_nw_n}\boldsymbol{G}_{nl}^{h_nw_n} \\ &\quad - \frac{1}{H^2W^2}\sum_{h_pw_p}\boldsymbol{S}_{apl}^{h_aw_a,h_pw_p}\boldsymbol{G}_{pl}^{h_pw_p}, \\ \frac{\partial\mathcal{L}}{\partial \boldsymbol{G}_{pl}^{h_pw_p}} &= -\frac{1}{H^2W^2}\sum_{h_aw_a}\boldsymbol{S}_{apl}^{h_aw_a,h_pw_p}\boldsymbol{G}_{al}^{h_aw_a}, \\ \frac{\partial\mathcal{L}}{\partial \boldsymbol{G}_{nl}^{h_nw_n}} &= \frac{1}{H^2W^2}\sum_{h_aw_a}\boldsymbol{S}_{anl}^{h_aw_a,h_nw_n}\boldsymbol{G}_{al}^{h_aw_a}.\end{aligned} \quad (20)$$

If $\boldsymbol{S}_{anl}^{h_aw_a,h_nw_n}$ and $\boldsymbol{S}_{apl}^{h_aw_a,h_pw_p}$ are viewed as two weighting factors for different positions, the gradient term with respect to $\boldsymbol{G}_{al}^{h_aw_a}$ is equivalent to the difference between the weighted aggregation of all $H \times W$ positive and negative part descriptors. In the same way, the gradient terms with respect to $\boldsymbol{G}_{pl}^{h_pw_p}$ and $\boldsymbol{G}_{nl}^{h_nw_n}$ are equivalent to the weighted summation of all $H \times W$ part descriptors of anchor samples.

When $n_o \geq 2$, $\boldsymbol{S}_{anl}^{h_aw_a,h_nw_n}$ and $\boldsymbol{S}_{apl}^{h_aw_a,h_pw_p}$ can be viewed as the high-order similarities. As the order n_o increases, the gradients of aligned part descriptors are highlighted over the misaligned parts. In this case, the gradient term $\partial\mathcal{L}/\partial\boldsymbol{G}_{al}^{h_aw_a}$ pushes the anchor descriptor $\boldsymbol{G}_{al}^{h_aw_a}$ close to the aligned positive part descriptors and away from the aligned negative part

TABLE I: Additional ablation study on HOReID with ResNet-50. "*w/.* 1×1" with the blue marker ✔ means using 1×1 kernels in the residual block of the HOReID; otherwise using 3×3 kernels. "*w/.* shared" with the blue marker ✔ represents that a single descriptor sampler of the LHP layer is shared across multiple levels; otherwise using unshared samplers for different feature levels. "*w/.* ReLU" with the blue marker ✔ represents that a ReLU layer is used after multilevel feature maps; otherwise we do not use ReLU. "*w/.* Hier." with the blue marker ✔ denotes the hierarchical structure is used in the HOReID framework; otherwise we use the non-hierarchical structure.

w/. 1×1	*w/.* shared	*w/.* ReLU	*w/.* Hier.	Market1501-SQ		Market1501-MQ		DukeMTMC		MSMT17	
				R1	mAP	R1	mAP	R1	mAP	R1	mAP
✘	✘	✘	✔	93.85	84.02	95.19	89.49	85.50	73.10	71.30	47.32
✔	✘	✘	✔	94.30	86.97	96.33	91.15	86.45	76.97	73.02	48.89
✘	✔	✘	✔	94.07	86.04	95.84	90.76	86.25	76.36	72.47	48.02
✔	✔	✔	✔	95.32	87.98	96.78	91.85	86.45	77.03	72.30	47.68
✔	✔	✘	✘	94.48	86.92	95.67	90.75	87.29	78.14	73.26	49.13
✔	✔	✘	✔	**95.74**	**88.78**	**96.97**	**92.36**	**88.12**	**79.79**	**74.37**	**50.39**

descriptors. Similarly, $\partial\mathcal{L}/\partial\boldsymbol{G}_{pl}^{h_p w_p}$ pushes the positive part descriptor $\boldsymbol{G}_{pl}^{h_p w_p}$ close to the aligned anchor part descriptors, while $\partial\mathcal{L}/\partial\boldsymbol{G}_{nl}^{h_n w_n}$ keeps the negative part descriptor $\boldsymbol{G}_{nl}^{h_n w_n}$ away from the aligned anchor part descriptors. When $n_o = 1$, we define $\boldsymbol{S}_{anl}^{h_a w_a, h_n w_n} = 1$ and $\boldsymbol{S}_{apl}^{h_a w_a, h_p w_p} = 1$, so the weighted summation is degraded into the averaged summation. That is, the gradient terms of the first-order model contain both aligned and misaligned part descriptors, which might generate even totally erroneous gradient directions. According to the above analysis, we treat the high-order weighting factor as a regularization term to regulate the gradient direction, which well explains 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 know the working principle of the proposed HOReID framework better. The HOReID framework not only provides regularization for the gradient direction but also enhances the pose-robustness property within the learned features.

E. General Attention Model

There are a few attention-based ReID works [5, 7, 31] which concentrate on refining semantic descriptors. In the present subsection, we will provide a detailed discussion on the differences between the HOReID and attention-based ReID models.

Given two input images $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$, we extract two convolutional feature maps $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{C \times H \times W}$ and their corresponding attention masks $\boldsymbol{M}_a, \boldsymbol{M}_b \in \mathbb{R}^{H \times W}$. Then, the two feature vectors are calculated by a GAP layer,

$$\begin{aligned} \boldsymbol{a} &= \frac{1}{HW} \sum_{h_a w_a} \boldsymbol{M}_a^{h_a w_a} \boldsymbol{A}^{h_a w_a} \\ \boldsymbol{b} &= \frac{1}{HW} \sum_{h_b w_b} \boldsymbol{M}_b^{h_b w_b} \boldsymbol{B}^{h_b w_b} \end{aligned} . \tag{21}$$

The similarity between $\boldsymbol{a}$ and $\boldsymbol{b}$ is formulated by,

$$\begin{aligned} \langle \boldsymbol{a}, \boldsymbol{b} \rangle &= \Big\langle \frac{1}{HW} \sum_{h_a w_a} \boldsymbol{M}_a^{h_a w_a} \boldsymbol{A}^{h_a w_a}, \frac{1}{HW} \sum_{h_b w_b} \boldsymbol{M}_b^{h_b w_b} \boldsymbol{B}^{h_b w_b} \Big\rangle \\ &= \frac{1}{H^2 W^2} \sum_{h_a w_a} \sum_{h_b w_b} \Big\langle \boldsymbol{M}_a^{h_a w_a} \boldsymbol{A}^{h_a w_a}, \boldsymbol{M}_b^{h_b w_b} \boldsymbol{B}^{h_b w_b} \Big\rangle \\ &= \frac{1}{H^2 W^2} \sum_{h_a w_a} \sum_{h_b w_b} \boldsymbol{M}_a^{h_a w_a} \boldsymbol{M}_b^{h_b w_b} \Big\langle \boldsymbol{A}^{h_a w_a}, \boldsymbol{B}^{h_b w_b} \Big\rangle \\ &= \frac{1}{H^2 W^2} \sum_{h_a w_a} \sum_{h_b w_b} \Big\langle \big[\boldsymbol{M}_a^{h_a w_a}\big], \big[\boldsymbol{M}_b^{h_b w_b}\big] \Big\rangle \Big\langle \boldsymbol{A}^{h_a w_a}, \boldsymbol{B}^{h_b w_b} \Big\rangle \end{aligned} , \tag{22}$$

where the local similarity $\langle \boldsymbol{A}^{h_a w_a}, \boldsymbol{B}^{h_b w_b} \rangle$ is weighted by the product of two attention values $\boldsymbol{M}_a^{h_a w_a} \boldsymbol{M}_b^{h_b w_b}$. Eq. 22 shows that the attention-based method can, to some extent, be seen as a second-order model. The attention product encodes the position relationships with two attention values in Eq. 22, while the GHP layer uses the inner product of two input vectors from different positions. That is, the GHP layer is a general attention model. Obviously, the inner product of two vectors is able to capture more complex and informative relationships of part pairs than the numerical product of two scalar values. Even though the attention model is well optimized, there is no guarantee that the attention product of the aligned part pair will be larger than the misaligned part one. 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 the similar values of attention product. In part, this explains why the GHP layer performs better than the standard attention model on resolving the pose misalignment problem.

VI. Experiments

A. Datasets

Market1501 [23]: 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 testing set, we select one image from each camera as a query image, forming 3,368 queries following the standard setting in [23]. We also adopt two evaluation protocols, *i.e.*, single-query (Market1501-SQ), which uses one image from each person, and multi-query (Market1501-MQ), which uses multiple images from each person. In the multi-query case, the features from multiple intra-class images captured by the same camera are averaged to produce a single descriptor.

CUHK03 [24]: It contains 14,097 images of 1,467 persons. It provides person bounding boxes detected both from the deformable part model detector and from manual labeling. We conduct experiments on the detected dataset and the original training/testing protocol is adopted following the previous work [24]. According to that protocol, 20 random splits are conducted on the dataset and average accuracy is reported.

DukeMTMC [25]: Similar to Market1501, DukeMTMC 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.

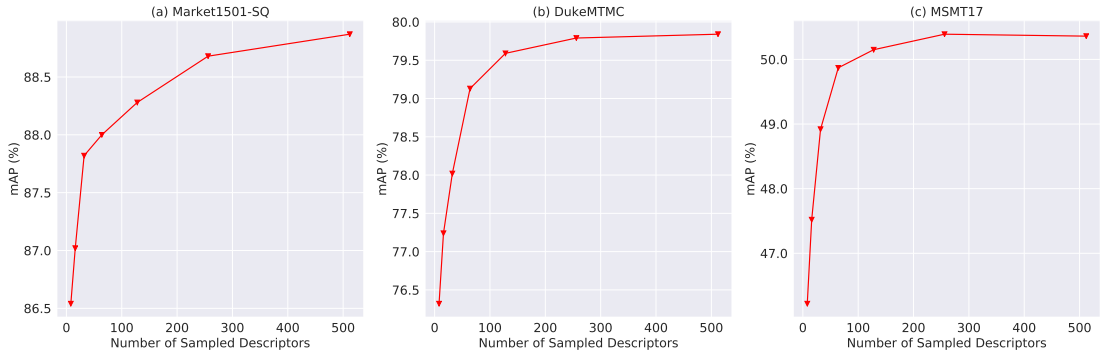


Fig. 8: Analysis on numbers of sampled descriptors with ResNet-50 on the Market1501, DukeMTMC and MSMT17 datasets.

MSMT17 [26]: 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 implement the proposed HOReID framework based on `Pytorch` [50]. We take the ResNet-50/101 [46] initialized with the parameters pretrained on ImageNet [51] as the backbone network. Following the work [14], the last fully-connected layer and global average pooling layer are removed and the stride of the last residual block *Conv4_1* is set from 2 to 1 for increasing the feature map size. Apart from the backbone network, the HOReID framework is lightweight, since it only adopts 1×1 convolutional kernels rather than 3×3 kernels to keep the receptive field of feature maps unchanged. Specifically, we set the dimension of global and part descriptors as $C_g = 512$ and $C_p = 512$. The number of sampled descriptors is set as $P = 256$. In order to reduce the computational complexity, the dimension of high-order features is decreased from 512^3 to 512, *i.e.*, $D_g = 512$ and $D_p = 512$.

Data Processing: In order to obtain enough context information from person images and a proper size of feature map for the proposed HOReID 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 [43], the training images are augmented with horizontal flipping and random erasing [52]. 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 [53], 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 64 and 4, respectively. Following [9], 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 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.

C. Ablation Study

Order of Aggregated Features: We first study the sensitivity of the order of aggregated features, which is associated with the number of hierarchical levels. Two interesting phenomena are observed in Fig. 7. First, a higher feature order benefits person ReID performance. The mAP scores of the three datasets increase with the feature order until they reach a stable performance. For example, the third-order GHP layer ($n = 3$) outperforms the first-order one ($n = 1$) by **6.27%**, **8.40%** and **10.01%** in terms of mAP on the three datasets respectively. Second, increasing the order ($n > 3$) makes a limited contribution to the mAP improvement compared with $n = 3$. To some extend, this is because the third-order model has largely eliminated person misalignments. Therefore, there is little room for further pose alignment improvements. We recommend $n = 3$ as it strikes a satisfactory balance between the computational efficiency and retrieval performance.

Joint GHP and LHP Layers: Then we investigate the contributions of the GHP and LHP layers on pose-invariant representation learning. In Fig. 7, the results show that the LHP layer consistently achieves superior mAP scores than the GHP layer. This phenomenon indicates that body-based features are more suitable than image-based features in pose 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 leveraging these two features significantly outperforms either of them on the three datasets. Accordingly, the body-based person representations have some advantages over the image-

TABLE II: Analysis of computational costs. "ResNet-50 + Cat + GAP" represents aggregating multi-level feature maps by the concatenation and *Global Average Pooling* (GAP), which can be viewed as the additional baseline model for "ResNet-50 + GHP". "ResNet-50 + Atten + Cat + GAP" denotes that local features are obtained by confidence attention maps and then both global and local features are concatenated for testing, which can be viewed as the additional baseline model for "ResNet-50 + GHP + LHP + DSR". The inference time and speed of all models are tested on a single GTX 1080 Ti GPU.

Method	#Level	#Param	FLOPS	Speed (fps)	Time (ms)	Market1501-SQ		DukeMTMC		MSMT17	
						R1	mAP	R1	mAP	R1	mAP
ResNet-50 + GAP	$n=1$	24.56M	6.28G	125.34	7.98	91.95	80.13	81.85	66.86	61.48	36.03
ResNet-50 + Cat + GAP	$n=2$	25.72M	6.39G	122.76	8.15	92.06	81.65	82.76	68.89	65.34	37.07
ResNet-50 + Atten + Cat + GAP	$n=2$	26.86M	6.50G	121.83	8.21	92.75	82.34	82.99	69.56	67.27	40.16
ResNet-50 + GHP	$n=2$	25.87M	6.43G	120.97	8.26	94.21	85.62	85.68	74.20	70.30	45.32
ResNet-50 + GHP + LHP + DSR	$n=2$	27.05M	6.56G	102.42	9.76	94.98	87.92	86.78	77.60	72.05	48.30
ResNet-50 + Cat + GAP	$n=3$	26.47M	6.46G	114.65	8.72	92.17	81.65	82.43	67.79	67.32	37.90
ResNet-50 + Atten + Cat + GAP	$n=3$	27.72M	6.62G	110.07	9.09	92.76	82.47	82.86	70.03	67.73	40.82
ResNet-50 + GHP	$n=3$	26.65M	6.53G	109.03	9.17	94.54	86.40	86.45	75.36	71.79	46.44
ResNet-50 + GHP + LHP + DSR	$n=3$	27.99M	6.76G	95.08	10.52	95.74	88.78	88.12	79.79	74.37	50.39

based ones, but they are complementary to each other. The results also show that sampling discriminative descriptors supervised by the proposed DSR is helpful for pose-robust feature learning. In the real-world applications, we would recommend to simultaneously learn both image-based and body-based features with the proposed DSR.

ReLU vs. No ReLU: As mentioned in Sec. III, we need to add a ReLU layer after multilevel feature maps $\{\boldsymbol{A}_l\}_{l=1}^n$ and $\{\boldsymbol{B}_l\}_{l=1}^n$ to make the part similarity be always non-negative in theory. Therefore, it is worthwhile to examine whether the HOReID framework can perform satisfactory pose alignment without ReLU. The results are reported in Table I. Interestingly, we observe that the HOReID framework without ReLU achieves superior performances than the one with ReLU. The main reason is that 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, which might hamper the representational capability of feature maps. Moreover, the experimental results partially reveal that the conclusion of Eq. 3 is also true for both positive and negative similarities. Therefore, in our architecture we do not use by default the ReLU layer after the multilevel feature maps.

Hierarchical vs. Non-Hierarchical Structure: In this part, we study the difference between the hierarchical and non-hierarchical structures. The hierarchical structure enriches the representational diversity of multiple hidden feature maps $\{\boldsymbol{X}_l\}_{l=1}^n$ by $(n-1)$ non-linearly residual blocks, *i.e.*, $\boldsymbol{X}_l \neq \boldsymbol{X}_{l'}, l \neq l'$. The non-hierarchical structure replaces the non-linearly residual block with an identity shortcut to enforce an equal feature map for multiple inputs, *i.e.*, $\boldsymbol{X}_l = \boldsymbol{X}_{l'}, l \neq l'$. From the results in Table I, it can be observed that the hierarchical structure performs better than the non-hierarchical structure on the three datasets. The reason is that the hierarchical structure brings more informative feature maps than the non-hierarchical one, resulting in a very strong high-order representational capability for the ReID models.

1×1 vs. 3×3 Kernels: Apart from the backbone network, we investigate the impact of different convolutional kernels (1×1 and 3×3) in the residual blocks of the proposed HOReID framework. From Table I, we can observe that using 1×1 convolutional kernels consistently outperforms 3×3 convolutional kernels with a large margin on all datasets. After analysis, we think the causes of this interesting phenomenon can be attributed into two sides. (1) The first reason is associated with the receptive fields of feature maps. 1×1 kernels keep the receptive fields unchanged for multilevel feature maps, while 3×3 kernels linearly enlarge the receptive fields as the increase of feature levels. Besides, feature maps with larger receptive fields usually contain less local pose information, which is unable to reduce pose misalignments. Therefore, 3×3 kernels perform inferior than 1×1 kernels for pose alignment. (2) The second reason is related to the interlevel inconsistency. As 3×3 kernels merge the 9 adjacent part descriptors into one part descriptor, the merged descriptor might include information from various semantic parts. For example, if a certain part descriptor only contains the semantic information of hands, the output part descriptor at the same position may be invaded by clothing and background due to 3×3 kernels. Worse still, the higher-level descriptors of hand parts are likely to be totally occupied by other body parts with larger areas. Since the high-order similarity is formed by the product of multilevel similarities at the same location, the similarity product between different semantic parts at the same position might degrade pose alignment. According to the analysis above, we recommend the 1×1 convolutional kernels to the HOReID framework instead of 3×3 kernels with the consideration of both computational efficiency and retrieve performance.

Shared vs. Unshared Samplers: In this part, we explore the effectiveness of the proposed shared sampler for semantic descriptors. Here we also design unshared samplers for different feature levels, and global features are input into their corresponding samplers to generate semantic part descriptors. From the results reported in Table I, we observe that the shared samplers achieve significant ReID performance improvements over the unshared ones on all datasets. The reason is that different samplers might exhibit different distributions of confidence maps, resulting in interlevel inconsistency. That is, the sampled descriptors of different feature levels at the same location represent different semantic information. In addition, the high-order feature aggregation requires the interlevel consistency of multilevel feature maps in order to guarantee the consistent weighting for aligned part pairs. In real-world applications, we recommend the shared sampler to the HOReID framework.

Number of Sampled Descriptors: Fig. 8 shows the ReID performance of the HOReID with different numbers of sam-

TABLE III: Accuracy (%) on the four ReID datasets. For a fair comparison, all reported results do not use the re-ranking proposed in [54]. Two baseline models ResNet-50/101 are trained with the triplet loss and leverage the global person features from the GAP layer to perform ReID evaluation.

Method	Market1501-SQ		Market1501-MQ		CUHK03		DukeMTMC		MSMT17	
	R1	mAP	R1	mAP	R1	R5	R1	mAP	R1	mAP
PDC [2]	84.14	63.41	-	-	78.29	94.83	-	-	58.0	29.7
GLAD [28]	89.9	73.9	-	-	82.2	95.8	-	-	61.4	34.0
KPM [55]	90.1	75.3	-	-	-	-	80.3	63.2	-	-
CRF [56]	93.5	81.6	-	-	88.8	97.2	84.9	69.5	-	-
HA-CNN [42]	91.2	75.7	93.8	82.8	-	-	80.5	63.8	-	-
PN-GAN [8]	89.43	72.58	92.93	80.19	79.76	96.24	73.58	53.20	-	-
Mancs [43]	93.1	82.3	95.4	87.5	92.4	98.8	84.9	71.8	-	-
PAB [9]	91.7	79.6	94.0	85.2	88.0	97.6	84.4	69.3	-	-
PCB + RPP [13]	93.8	81.6	-	-	-	-	83.3	69.2	68.2	40.4
SGGNN [57]	92.3	82.8	-	-	-	-	81.1	68.2	-	-
MGN [15]	95.7	86.9	96.9	90.7	-	-	88.7	78.4	-	-
FD-GAN [10]	90.5	77.7	-	-	-	-	80.0	64.5	-	-
HPM [14]	94.2	82.7	-	-	-	-	86.6	74.3	-	-
RollBack [58]	92.5	79.9	-	-	-	-	85.2	70.2	-	-
LITM + GHIS [59]	93.9	83.9	-	-	-	-	85.9	74.5	-	-
DSA [11]	95.7	87.6	-	-	-	-	86.2	74.3	-	-
CASN [60]	94.4	82.8	-	-	-	-	87.7	73.7	-	-
VPM [61]	93.0	80.8	-	-	-	-	83.6	72.6	-	-
ResNet-50	91.95	80.13	93.67	85.34	76.50	93.47	81.85	66.86	61.48	36.03
ResNet-50 + HOReID	95.74	88.78	96.97	92.36	94.02	99.47	88.12	79.79	74.37	50.39
ResNet-50 + HOReID + GHIS [59]	96.45	90.00	97.57	93.14	95.66	99.63	89.12	81.03	76.24	52.97
ResNet-101	93.10	82.57	94.09	86.44	79.18	94.58	82.83	67.96	63.93	40.13
ResNet-101 + HOReID	96.27	90.06	97.23	93.18	95.76	99.62	88.95	81.93	76.89	53.24
ResNet-101 + HOReID + GHIS [59]	**96.79**	**91.04**	**97.78**	**93.94**	**96.12**	**99.71**	**89.83**	**82.16**	**78.42**	**54.77**

pled descriptors, *e.g.*, $P \in \{8, 16, 32, 64, 128, 256, 512\}$. As seen, the HOReID reaches the best performance with 256 sampled descriptors. Intuitively, the number P of sampled descriptors determines the diversity of the sampled part descriptors. When $P = 0$, the resulting system is equivalent to the HOReID framework with only the GHP layer. With the increasing of the number P, the mAP scores are significantly improved by **2.14%** on Market1501, **3.47%** on DukeMTMC and **3.33%** on MSMT17 from $P = 8$ to $P = 256$. This indicates that enlarging the number of sampled descriptors may contribute to the discriminative ability of sampled body-based features. We also try more dense sampling settings, such as 512. However, a greater number of sampled descriptors will incur additional computational cost without leading to any observable performance improvement. Therefore, we adopt $P = 256$ for sampling descriptors in this work.

Analysis on Computational Costs: In Table II, we further analyze the computational costs of the proposed HOReID framework. As shown in Fig. 3, we adopt the two different strategies in order to maintain high computation efficiency with the comparison to the baseline model. For reducing the number of parameters of GHP and LHP layers, we use two residual blocks with a 1×1 convolutional layer, a BatchNorm layer and a ReLU layer to extract multilevel feature maps. Besides, two off-the-shelf GPU-based `Pytorch` functions, *i.e.*, `torch.fft` and `torch.ifft`, are applied to implement FFT and IFFT on GPUs, making high-order pooling more computationally efficient and effective. Compared with our baseline "ResNet-50 + GAP", "ResNet-50 + GHP" has achieved superior performance and only brings additional 1M $\sim$ 2M network parameters. Compared with "ResNet-50 + Cat + GAP", "ResNet-50 + GHP" has similar computational cost but achieves higher accuracies, which shows that the improvements of the GHP layers are not brought by the additional computational costs. Besides, we also observe that "ResNet-50 + GHP + LHP + DSR" performs better than "ResNet-50 + Atten + Cat + GAP" on the condition of the similar model parameters. More interestingly, the running speed of the HOReID is more than 95 FPS, which is fast enough for video-based person Re-ID in real-time.

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

Finally, we report the performance of the HOReID and other state-of-the-art methods in Table III. To avoid a risk of the local optima, we adopt a *Global Hard Identity Searching* (GHIS) [59] method to adaptively select hard classes and make the training more efficient and effective. By fine-tuning the framework based on ResNet-101, the HOReID clearly attains the best performance on each dataset. Compared with the pose-based method like DSA [11], the HOReID achieves significant gains of **3.44%** and **7.86%** for mAP on Market1501 and DukeMTMC, indicating the advantage of the proposed HOReID. Moreover, our framework is more efficient: the HOReID performs superior pose alignment with only identity labels, while other systems require human pose information or person 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 HOReID is the best-performing method, outperforming the second best method by **14.37%** for mAP. This clearly demonstrates that the HOReID achieves a satisfactory generalization on the large-scale dataset.

VII. Conclusion

In this paper, we propose a novel framework, named HOReID, which intends to address the ubiquitous pose misalignment problem. Two novel GHP and LHP layers are proposed to capture global and local high-order interactions to highlight aligned similarities and reduce misaligned similarities. Additionally, we put forward a shared descriptor sampler regulated by the novel DSR to enhance the diversity of sampled descriptors for all feature levels. Due to the high-order mapping, it becomes possible to conduct the fine-grained pose alignment for the person images without relying on human pose information. In the future, we will extend this work to the fields of attribute recognition, face recognition and vehicle re-identification, where the pose misalignment problem is prevalent.

References

[1] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1077–1085.

[2] 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.

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

[4] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4099–4108.

[5] 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.

[6] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 420–429.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[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] H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang, "Alignedreid++: Dynamically matching local information for person re-identification," in *Pattern Recognition*, vol. 94. Elsevier, 2019, pp. 53–61.

[13] 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.

[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] 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.

[16] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 384–393.

[17] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10. IEEE, 2018, pp. 3037–3045.

[18] Y. Guo and N.-M. Cheung, "Efficient and deep person re-identification using multi-level similarity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2335–2344.

[19] H. Luo, W. Jiang, X. Fan, and C. Zhang, "Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification," *IEEE Transactions on Multimedia*, 2020.

[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*. Springer, 2016, pp. 34–50.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," in *IEEE Transactions on Image Processing*, vol. 28, no. 9. IEEE, 2019, pp. 4500–4509.

[30] 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.
[31] 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.
[32] Y. Hao, N. Wang, X. Gao, J. Li, and X. Wang, "Dual-alignment feature embedding for cross-modality person re-identification," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 57–65.
[33] P. Wang, F. Su, Z. Zhao, Y. Zhao, L. Yang, and Y. Li, "Deep hard modality alignment for visible thermal person re-identification," in *Pattern Recognition Letters*, vol. 133. Elsevier, 2020, pp. 195–201.
[34] 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," in *IEEE Transactions on Multimedia*. IEEE, 2020.
[35] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in neural information processing systems*, 2015, pp. 2017–2025.
[36] 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.
[37] 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.
[38] 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 *International Conference on Learning Representations*, 2017.
[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] S. Cai, W. Zuo, and L. Zhang, "Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 511–520.
[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] 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.
[43] 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.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
[45] 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.
[46] 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.
[47] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.
[48] 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.
[49] R. Pagh, "Compressed matrix multiplication," in *ACM Transactions on Computation Theory (TOCT)*, vol. 5, no. 3. ACM New York, NY, USA, 2013, pp. 1–17.
[50] 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.
[51] 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.
[52] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation." in *AAAI*, 2020, pp. 13 001–13 008.
[53] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," in *arXiv preprint arXiv:1703.07737*, 2017.
[54] 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.
[55] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6886–6895.
[56] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang, "Group consistent similarity learning via deep crf for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8649–8658.
[57] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 486–504.
[58] Y. Ro, J. Choi, D. U. Jo, B. Heo, J. Lim, and J. Y. Choi, "Backbone cannot be trained at once: Rolling back to pre-trained network for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8859–8867.
[59] Y. Zhang, Q. Zhong, L. Ma, D. Xie, and S. Pu, "Learning incremental triplet margin for person re-identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 9243–9250.
[60] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5735–5744.
[61] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 393–402.

Pingyu Wang is currently a Ph.D. candidate at the Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include attribute classification, face recognition, person reidentification and deep learning.

Zhicheng Zhao is an associate professor of Beijing University of Posts and Telecommunications. He was a visiting scholar at School of Computer Science, Carnegie Mellon University from 2015 to 2016. His research interests are computer vision, image and video semantic understanding and retrieval. He has authored and coauthored more than 60 journal and conference papers.

Fei Su is a female professor in the multimedia communication and pattern recognition lab, Beijing university of posts and telecommunications. She received the Ph.D. degree majoring in Communication and Electrical Systems from BUPT in 2000. She was a visiting scholar at electrical computer engineering department, Carnegie Mellon University from 2008 to 2009, Her current interests include pattern recognition, image and video processing and biometrics. She has authored and co-authored more than 70 journal and conference papers and some textbooks.

Xingyu Zu is currently a master candidate at the Beijing Key Laboratory of Network System and Network Culture, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include face recognition and deep learning.

Nikolaos V. Boulgouris is with the Department of Electronic and Computer Engineering of Brunel University, London, U.K. From 2004 to 2010 he was an academic member of staff with King's College London, and prior to that he was a researcher with the Department of Electrical and Computer Engineering of the University of Toronto, Canada. Dr. Boulgouris served as Technical Program Chair for the 2018 IEEE International Conference on Image Processing (ICIP). He has served as Senior Area Editor for the IEEE Transactions on Image Processing and as Associate Editor for the IEEE Transactions on Circuits and Systems for Video Technology, from which he received the 2017 Best Associate Editor Award. In the past he served as Associate Editor for the IEEE Transactions on Image Processing and for the IEEE Signal Processing Letters. He was co-editor of the book Biometrics: Theory, Methods, and Applications, which was published by Wiley - IEEE Press, and guest co-editor for two journal special issues. From 2014 to 2019 he served as member of the IEEE Image, Video, and Multidimensional Signal Processing Technical Committee (IVMSP - TC). Dr. Boulgouris is a Senior Member of the IEEE and a Fellow of the Higher Education Academy.