Joint uneven channel information network with blend metric loss for person re-identification

Zhi Yu · Zhiyong Huang · Wencheng Qin · Tianhui Guan · Yuanhong Zhong · Daming Sun

Received: 2 August 2021 / Accepted: 27 February 2022 / Published online: 22 March 2022
© The Author(s) 2022

Abstract
Person re-identification, one of the most challenging tasks in the field of computer vision, aims to recognize the same person cross different cameras. The local feature information has been proved that can improve performance efficiently. Image horizontal even division and pose estimation are two popular methods to extract the local feature. However, the former may cause misalignment, the latter needs much calculation. To fill this gap and improve performance, an efficient strategy is proposed in this work. First, a joint uneven channel information network consisting of an uneven channel information extraction network and a channel information fusion network is designed. Different from the traditional image division, the former can divide images horizontally and unevenly with strong alignment based on weak pose estimation, and extract multiple channel information. The latter can joint channel information based on channel validity and generate an efficient similarity descriptor. To optimize the joint uneven channel information network efficiently, this work proposes a blend metric loss. The extra image information is utilized to dynamically adjust the penalty for the sample distance and the distance margin based on the outlier of the hardest sample to construct i-TriHard loss. Besides, softmax loss and center loss are embedded in the blend metric loss, which can guide the network to learn more discriminative features. Our method achieves 89.6% mAP and 95.9% Rank-1 on Market-1501, 79.9% mAP and 89.4% Rank-1 on DukeMTMC. The proposed method also performs excellently on occluded datasets.

Keywords Person re-identification · Joint uneven channel information · Blend metric loss · Dynamic penalty

Introduction
With the popularity of video surveillance, person re-identification technology has attracted more and more attention. Given an image of a person taken by a specific camera, it can retrieve other images of this person from the gallery taken by other cameras. With the wide application in practical monitoring and intelligent security fields, it has always been a mainstream subject.

Traditional methods use manually designed features, which have achieved results, but they are still far from being applied in actual scenes. With the quick development of deep learning, person re-identification has also made rapid progress. Studies of person re-identification based on deep learning can be divided into two branches: representation learning and metric learning. Representation learning does not directly consider the similarity between images when training the network, but treats the person re-identification as a classification task or a verification task. Metric learning aims to learn the similarity of two images through the network. In the person re-identification, the similarity between different images from the same person is greater than that of different images from different persons.

Most of researches concentrate on extracting the global feature [1,2]. Newer researches attempt to divide the image into small pieces [3] and utilizes pose estimation to extract different body regions [4,5]. However, the former may cause misalignment, the latter needs much calculation. To fill this gap, a joint uneven channel information network (JUCIN) consisting of an uneven channel information extraction network and a channel information fusion network is proposed. The former, different from the traditional horizontal even division, a weak pose estimation module combining the advantages of even division and pose estimation is utilized to divide image into four uneven local channels with strong
alignment, and consider the original image as a global channel. Therefore, five channel information can be generated. Next, the latter fuses five pairs of channel information into the similarity descriptor, and a dynamic fusion strategy based on channel validity is embedded in the pipeline. Last, this descriptor is used to finish person re-identification task.

Contrastive loss [6], triplet loss [7], TriHard loss [8], quadruplet loss [9], and margin sample mining loss [10] are proposed to optimize the neural network. It is different from representation learning in that metric learning relies on sample pairs. In this work, to optimize the feature extraction networks in JUCIN, a blend metric loss (BML) is proposed. First, TriHard loss is improved to construct i-TriHard loss, which can utilize the extra image information to dynamically adjust the penalty for the sample distance and the distance margin, thereby optimizing the spatial distribution between positive samples and negative samples. Besides, softmax loss and center loss [11] are embedded in the blend metric loss, which can guide the network to learn more discriminate features.

The proposed method shows excellent performance when compared to state-of-the-art methods. The contributions of this work can be summarized as follows: (1) A weak pose estimation strategy is proposed to guide horizontal uneven division, this allows the module to achieve channel alignment efficiently. (2) A joint uneven channel information network based on weak pose estimation is proposed to extract and fuse channel information dynamically. (3) A blend metric loss is proposed to optimize network, which can optimize the spatial distribution between samples to enhance the performance.

Related work

In the related researches of representation learning, He et al. [12] used a convolutional neural network to extract the global feature. Additionally, Lin et al. [13] introduced different attribute features such as hair and clothes to improve generalization. Vairio et al. [14] divided the images into several pieces, and then input these pieces into a long short-term memory network to generate the final features. However, this method requires images had strong alignment. Zhao et al. [15] proposed a Spindle Net, utilizing a convolutional pose machines to locate landmarks and extract different body regions. Then, it could fuse the global features and the local features. Zheng et al. [16] used a pose estimation network and an affined transformation to divide the image into several areas, and then used a PoseBox to correct the local areas. The local areas and the original image were then input into the network to extract the features. Wei et al. [17] introduced the global-locallignment descriptor to reduce the negative impact of the changeable pose. The image was divided into three areas, and the global average pooling method could solved the mismatch of the image. The fusion feature was then extracted. Zheng et al. [18] designed a SP distance adaptive alignment model, which could align local features without additional information, and used a dynamic alignment algorithm to find the shortest path. Yu et al. [19] proposed a deep discriminative representation method to learn the features and impose a discriminative constraint on the feature representation.

In the related researches of metric learning, Varior et al. [6] proposed a contrast loss and used it in the twin network, which could dynamically adjust the distance threshold according. Schroff et al. [7] proposed a triplet loss. The network took the triple image as input, which could pull the distance between positive sample pairs and push the distance between negative sample pairs, efficiently improving feature discrimination and re-identification performance. Hermans et al. [8] proposed a TriHard loss. For each sample batch, the loss selected a positive sample and a negative sample that were the most difficult to distinguish, and then constructed a sample set to calculate the final loss. Chen et al. [9] proposed a quadruple loss, which took four different images as the input of the network. Compared to triplet loss, quadruple loss considered the absolute distance between positive samples and negative samples. Zhu et al. [20] designed a hybrid similarity function to measure the similarity between the feature pairs, they also proposed a deep hybrid similarity learning method, which could reasonably allocate the complexity of feature learning and measurement learning, thereby improving the performance. Liu et al. [21] proposed a nonlinear deep metric learning strategy based on deep belief networks and component analysis, which used a data change methods to maximize the number of images and learned nonlinear feature mapping. Din et al. [22] improved the loss function and optimized the learning algorithm of network. They proposed a deep neural network with scalable distance-driven feature learning, which could effectively improve the performance.

Joint uneven channel information network

A joint uneven channel information network (JUCIN) is proposed in Fig. 1, which consists of an uneven channel information extract network and a channel information fusion network. First, the image is input to a weak pose estimation module, which can locate five landmarks and divide the image into four different pieces unevenly. Different from the traditional horizontal division, this strategy can achieve image division with alignment. Different pieces are defined as different channels in this work, and then utilize convolutional neural networks to extract the channel information. To optimize the feature extraction networks, a blend metric loss is proposed, the detailed information will be explained in Sect. 4. A channel information fusion network is proposed
Fig. 1 The pipeline of JUCIN. It consists of an uneven channel information extract networks (UCIEN) and a channel information fusion networks (CIFN). WPEM refers to weak pose estimation module. GCI refers to global channel information. LCI refers to local channel information. BML refers to blend metric loss. WDM refers to weight decision module. IFM refers to information fusion module.

to fuse different channel information from two compared images. Different channel information should contribute differently. Therefore, fusion weight which is generated by a weight decision module based on channel validity is introduced. Last, the weight and the channel information are combined to generate the final similarity descriptor for person re-identification.

Uneven channel information extraction network

The uneven channel information extraction network can extract a global channel information and four uneven local channel information. As the core of the extraction network, the weak pose estimation module can divide the image horizontally and unevenly with strong alignment. The details are illustrated in the following sections.

Weak pose estimation

As mentioned above, the local feature information can improve the performance. The popular method to extract the local feature is dividing images into several pieces evenly or generating some local regions based on pose estimation. However, the former requires the image pair to have strong alignment, or else this method will cause extra negative influence due to misalignment as shown in Fig. 2. The latter can solve misalignment but needs much additional computation.

To balance these two methods and improve the traditional image division method, a novel division strategy which use a weak pose estimation module is proposed. It can combine the advantages from horizontal even division and pose estimation, achieving unevenly but efficient image division without much calculation.

The pipeline of weak pose estimation module is shown in Fig. 3. For an input image, it first detects five landmarks: top of head, neck, crotch, left foot, and right foot, and then divides the image into four local channels horizontally and evenly.

Inspired by CPM [23], the weak pose estimation module utilizes a similar mechanism to locate different landmarks. But in this work, five landmarks are enough to divide the images. Therefore, some stages and neural layers are reduced.

After determining five landmarks, the following equations are utilized to generate four local channels:

\[
\begin{align*}
    c_1 &= \text{input}[\text{width}, y(L_1) : \max(y(L_4), y(L_5))], \\
    c_2 &= \text{input}[\text{width}, y(L_1) : y(L_2)], \\
    c_3 &= \text{input}[\text{width}, y(L_2) : y(L_3)], \\
    c_4 &= \text{input}[\text{width}, y(L_3) : \max(y(L_4), y(L_5))].
\end{align*}
\]
where input is the matrix of the input image, width is the image width, \( L_i \) is the \( i \)th landmark, and \( y(L_i) \) is the height coordinate of the landmark.

In summary, the weak pose estimation module utilizes \( L_1, L_4, \) and \( L_5 \) to generate \( c_1 \), utilizes \( L_1 \) and \( L_2 \) to generate \( c_2 \), utilizes \( L_2 \) and \( L_3 \) to generate \( c_3 \), and utilizes \( L_3, L_4, \) and \( L_5 \) to generate \( c_4 \). Based on this, the input image can be divided horizontally and unevenly into four different channels, and make these channels as local channels. The traditional division method may cause misalignment, but this novel method combines the simplicity of horizontal division and the efficiency of pose estimation. It can accomplish the image division with less computation and resolve the misalignment between the image pairs.

Uneven channel information extraction

In this work, an uneven channel information extraction network is proposed to extract the channel information. As illustrated in Fig. 1, it consists of five branches.

The first branch takes the original image as input, which is used to extract the global channel information. The second branch takes the processed image consisting of three pieces as input, which can reduce the negative influence of background and extract the local channel information. The last three branches take three horizontally divided images as input, which are all part of the original image and do not overlap. Finally, three local channel information can be generated.

The pipeline of information extraction in each branch is shown in Fig. 4, and the core module is ResNet50. The batch normalization (BN) layer can constrain the distribution of the features and balance the dimensions of different features.

Therefore, for an image, five different channel information can be extracted, including a global channel information and four uneven local channel information.

For the image pair to be compared, the weak pose estimation module gives the corresponding channel strong alignment, which can solve the misalignment problem between image pieces without increasing much calculation. the uneven channel information extraction network can also make the channel information more robust.

In summary, the uneven channel information extraction network based on a weak pose estimation combines image horizontal division and pose estimation, which can achieve efficient alignment between image pieces. It aims to learn a global channel information and four local channel information. The pipeline can be divided into five branches in structure and function. An input image can generate five feature information through this pipeline. This can be summarized as follows:

\[
f^{set} = [f^{c_0}, f^{c_1}, f^{c_2}, f^{c_3}, f^{c_4}],
\]

where \( f^{c_0} \) is the global channel information, \( f^{c_1}, f^{c_2}, f^{c_3}, f^{c_4} \) are four local channel information.

Channel information fusion network

Five channel information can be obtained through the uneven channel information extraction network. Dealing with this information and combining them into an efficient descriptor is extremely important. A channel information fusion network is proposed to accomplish this task. As shown in Fig. 1, it consists of a weight decision module and an information fusion module. The former is the core of the fusion network and can generate different fusion weights based on channel validity. The latter can fuse the channel information similarity with the fusion weight to generate the final similarity descriptor.

For an image pair consisting of a probe image and a gallery image, it can obtain five channel information pairs. Then, handling these information pairs to generate a similarity descriptor should be focused. In this work, different contributions from channels to final similarity should be weighted. To accomplish this, a weight decision module based on channel validity is embedded to generate the channel weight dynamically. Finally, the channel information fusion module fuses dynamic weights and channel information into the similarity descriptor. The similarity between the probe image and the gallery image can be presented as follows:

\[
S = \sum_{i=0}^{N-1} w_i \ast s(f^{c_i}_p, f^{c_i}_g),
\]

where \( N \) is the number of channels, \( w_i \) is the weight of \( i \)th channel, \( s(\cdot) \) stands for the similarity between the \( i \)th channel information from the probe image and the gallery image. \( f^{c_i}_p \) and \( f^{c_i}_g \) are the \( i \)th channel information from the probe image and the gallery image.

The core of the channel information fusion network is determining the weight from different channels. In this pipeline, a weight decision module is designed. The following section will introduce the working mechanism.

The local channel information is considered as the supplementary information for the global channel information in this work, the former is more important than other channel information. Because there are some deviations in the
Fig. 5 An example of channel information fusion. Assume the global channel information similarity is $s = 0.7$, two channel validities are $v_p = 1.0$ and $v_g = 1.0$, and the kernel function is average. Therefore, the weight of global channel information is $w = 1.0$. The similarity and weight for other channels are calculated by the same way. This work uses the weighted sum to generate the final similarity.

Figure 5 shows an example of channel information fusion. Five pairs of channel similarity information and channel weights are combined to generate the final similarity. The fusion similarity generated by the information fusion module is considered the discriminative similarity descriptor for person re-identification in this work. A higher fusion similarity indicates that two compared images have a higher similarity.

Blend metric loss
To optimize the feature extraction networks efficiently in JUCIN, a blend metric loss (BML) is proposed in Fig. 6, which consists of the improved i-TriHard loss, softmax loss, and center loss. The dynamic penalty is imposed on TriHard loss based on extra image information to construct i-TriHard loss, which is the core of BML and can optimize the spatial distribution of the samples. The strategy for dynamic distance margin adjustment is also embedded in this pipeline. BML can guide the network to learn more discriminate features in the embedding space. The following sections show the details.

Improved i-TriHard loss
Triplet loss [7] is widely used in metric learning, and a large number of subsequent losses are derived from that, which can guide the network to learn more discriminative features and improve robustness. Triplet loss takes a positive sample pair and a negative sample pair as input, which are $a$ (anchor), $p$ (positive sample), and $n$ (negative sample). The positive sample pair is $a$ and $p$, while the negative sample pair is $a$ and $n$. Triplet loss aims to maximize the distance between
the negative samples and minimize the distance between the positive samples. Triplet loss is formulated as follows:

$$L_{\text{triplet}} = \frac{1}{N} \sum_{i=1}^{N} [d(a_i, p_i) - d(a_i, n_i) + m]^+, \quad (9)$$

where $d(\cdot)$ is Euclidean distance, $a_i$ is the $i$th anchor, $p_i$ is the $i$th positive sample, $n_i$ is the $i$th negative sample, and $m$ is the distance margin.

As an evolution of triplet loss, TriHard loss [8] assumes the positive samples and the negative samples in triplet loss are easy to distinguish, which is not conducive to network training. By selecting the farthest positive sample and the nearest negative sample to optimize the network, TriHard loss can enhance the generalization and the representation of the network. TriHard loss is formulated as follows:

$$L_{\text{TriHard}} = \frac{1}{P \times K} \sum_{i=1}^{P \times K} [\max(d(a_i, p_i)) - \min(d(a_i, n_i)) + m]^+, \quad (10)$$

where $P$ is the number of identities, $K$ is the number of image from each identity.

TriHard loss uses the hardest positive sample and the hardest negative sample exclusively, assigning the same weight to the hardest sample for each anchor. However, it ignores the remaining positive samples and negative samples, and it cannot consider the specific distribution of the samples in the training batch.

Ideally, if the performance of the network is efficient, all positive samples or negative samples should be clustered together. If there is a large gap between the hardest positive sample and other positive samples, or there is a large gap between the hardest negative sample and other negative samples, it means that the penalty of the network needs to be further strengthened. Therefore, TriHard loss is improved to construct i-TriHard loss which can dynamically adjust the penalty for sample distance and distance margin. A specific description is illustrated in Fig. 7. Different penalties are formulated based on the distribution of the samples.
where \( P(a) \) and \( N(a) \) are all positive samples and all negative samples in a batch, \( N_p \) and \( N_n \) is the number of the positive samples and the negative samples in a batch.

By these definitions, \( w_p \) always satisfies \( w_p \geq 1 \) and \( w_n \) always satisfies \( w_n \leq 1 \). These weights can improve efficiency in early stages of training. As the training progresses, \( w_p \) slowly decreases and \( w_p \) slowly increases, which can gradually lower the penalty and it is conducive to the follow-up training.

The proposed i-TriHard loss is always greater than the original TriHard loss because of the new dynamic weights for the hardest positive sample distance and the hardest negative sample distance. If the loss remains a large value all the time, it will make the network training more difficult and result in underfitting. To solve this problem, the distance margin is embedded in the blend metric loss, which can appropriately reduce the distance threshold and the loss. The weight of this distance margin can be formulated as follows:

\[
w_m = \frac{w_n}{w_p}
\]

Using Eqs. (12), (13), and (14), Eq. (11) can be transformed as follows:

\[
L_{\text{i-TriHard}} = \frac{1}{P \times K} \sum_{i=1}^{P \times K} \left[ \frac{\max(d(a_i, p_i))^2}{\sum_{x \in P(a_i)} d(a_i, x)} + \frac{\min(d(a_i, n_i))^2}{\sum_{x \in N(a_i)} d(a_i, x)} + \frac{w_m}{w_p} \right].
\]

**Blend metric loss**

Softmax loss is the most popular classification loss which aims to construct several hyperplanes for dividing embedding space. It can be formulated as follows:

\[
L_{\text{softmax}} = -\sum_i y_i \log s_i,
\]

where \( y_i = 1 \) when the label of ith sample is the true label, otherwise, \( y_i = 0 \). \( s_i \) is the output of softmax.

As triplet loss cannot achieve optimal constraint at the global level, the combination with softmax loss can guide neural network to learn more discriminative features.

Besides, triplet loss and its variants only consider the distance difference between the hardest positive sample and the hardest negative sample. Although the discreteness between different classes is considered, it ignores the absolute value of the distance and the feature cohesion. Center loss [11] can solve this problem efficiently, it learns the center of each class and penalizes the loss through the distance between the features and the class center. Center loss is formulated as follows:

\[
L_{\text{center}} = \frac{1}{2} \sum_{i=1}^{m} d(x_i - c_{y_i}),
\]

where \( m \) is the batch size, \( x_i \) is the feature corresponding to the ith sample, \( c_{y_i} \) represents the feature center of the class \( y_i \).

In this work, i-TriHard loss, softmax loss, and center loss are combined to construct the blend metric loss, which can be formulated as follows:

\[
L_{\text{BML}} = L_{\text{i-TriHard}} + L_{\text{softmax}} + L_{\text{center}}.
\]

The dynamic adjustment for the hardest sample distance and the distance margin is embedded in the blend metric loss, and the combination of different loss is utilized to optimize the network, which can guide the network to learn more discriminative features.

**Experiments**

**Datasets**

Market-1501 [24] is collected from Tsinghua University, using six cameras to collect 32,668 images from 1501 different persons. The training set consists of 12,936 images from 751 persons, and the testing set consists of a query set containing 3368 images and a gallery set containing 19,732 images. DukeMTMC [25] is collected from Duke University, using eight cameras to collect 36,411 images from 1404 different persons. The training set consists of 16,522 images from 702 persons, and the testing set consists of a query set containing 3228 images and a gallery set containing 17,661 images. Occluded-DukeMTMC [26] is an occluded dataset, using eight cameras to collect 35,489 images from 1404 different persons. The training set consists of 15,618 images from 702 persons, and the testing set consists of a query set containing 2228 images and a gallery set containing 17,661 images. Partial iLIDS [27] is a simulated partial dataset, including 238 images from 119 person identities. Each label contains an occluded image and a non-occluded image.

As the most popular evaluation metrics, cumulative matching characteristic (CMC) curves and mean average precision (mAP) are utilized to evaluate the performance of person re-identification model in this work.
Implementation details

The weak pose estimation module is trained based on MPII dataset [28] to generate five landmarks and four local channels. ResNet50 is the core module in the uneven channel information extraction network, and it is initialized with ImageNet [29] pre-trained model. The input image is resized to 256 × 128. The batch size is set to 64, the number of different identities is set to 16, and the number of different images from per identity is set to 4. The maximum training epoch is set to 200. To prevent the overfitting, this work uses four different regularization strategies. First, the warmup learning rate is applied in our pipeline. The initial learning rate is $3 \times 10^{-6}$, and then rises to $3 \times 10^{-4}$, next drops to $3 \times 10^{-5}$ in the 50th epoch, and last drops to $3 \times 10^{-6}$ in the 100th epoch. Second, the random flipping and the random erasing strategy [43] are used to augmented the training data in our experiment. Third, dropout strategy is applied to inactivate certain nerve units, the rate of dropout is set as 0.8. Fourth, L2 regularization is used in the reverse optimization of the network. The whole experiment is implemented on the hardware equipped with a GeForce GTX 2080Ti GPU.

Comparison with state-of-the-arts

The proposed method is compared with state-of-the-arts in Table 1. Our method achieves 89.6% mAP and 95.9% Rank-1 in Market-1501, 79.9% mAP and 89.4% Rank-1 in DukeMTMC. Compared to the best method in the table, our method has different improvement on two datasets.

| Method       | Market-1501 mAP | Market-1501 Rank-1 | DukeMTMC mAP | DukeMTMC Rank-1 |
|--------------|-----------------|--------------------|--------------|-----------------|
| AWTL [31]    | 75.7            | 89.5               | 63.4         | 79.8            |
| Aligned [18] | 77.7            | 90.6               | 67.4         | 81.2            |
| PCB [32]     | 81.6            | 93.8               | 69.2         | 83.3            |
| BFE [33]     | 85.0            | 94.5               | 75.8         | 88.7            |
| AANet [34]   | 82.5            | 93.9               | 72.6         | 86.4            |
| DeepRSAM [35]| 78.5            | 94.2               | 64.8         | 84.2            |
| BoT [36]     | 85.9            | 94.5               | 76.4         | 86.4            |
| SNR [37]     | 84.7            | 94.4               | 72.9         | 84.4            |
| HO [38]      | 84.9            | 94.2               | 75.6         | 86.9            |
| DLBC [30]    | 87.4            | 94.6               | 78.5         | 88.7            |
| CDN [39]     | 86.0            | 95.1               | 76.8         | 88.6            |
| LDS [40]     | 87.3            | 95.0               | 76.1         | 88.2            |
| PAT [41]     | 88.0            | 95.4               | 78.2         | 88.8            |
| GPS [42]     | 87.8            | 95.2               | 78.7         | 88.2            |
| JUCIN (Ours) | 89.6            | 95.9               | 79.9         | 89.4            |

To further verify the effectiveness of our method, the occluded environment is also considered. The performance on Occluded-DukeMTMC and Partial-iLIDS datasets are shown in Tables 2 and 3.

In Table 2, our method achieves significant improvement compared to the best method HG [49], mAP is improved from 50.5 to 53.4% and Rank-1 is improved from 61.7 to 62.1%. In Table 3, our method has certain improvement compared to the best method LKWS [53], showing increase from 80.7 to 82.4% in Rank-1 and increase from 88.2 to 89.9% in Rank-3. From the data gathered, it can be concluded that our proposed method achieves advanced performance whether under a normal environments or an occluded environments. The data supports the effectiveness of the proposed method.

Ablation evaluation on channel information fusion

The proposed JUCIN framework can extract and joint multiple channel information. To prove the validity of the proposed fusion strategy, an ablation experiment based on different
channel information is conducted. The result is illustrated in Fig. 8. “Global” represents the global channel information. “Local_1”, “Local_2”, “Local_3”, and “Local_4” represent four local channel information.

Compared to the global channel information, our method obtains 1.2% mAP and 0.9% Rank-1 gains on Market-1501. It also obtains 1.4% mAP and 1% Rank-1 gains on DukeMTMC. The result can be improved further with the introduction of more additional channel information. When all channel information is merged, further improved results of 89.6% mAP and 95.9% Rank-1 can be obtained in Market-1501, 79.9% mAP and 89.4% Rank-1 in Duke-MTMC. The data demonstrate the effectiveness of our proposed method of channel information fusion.

**Evaluation on uneven division**

Different from the traditional even division, the uneven division combing pose estimation is utilized in this work. To prove the effectiveness, a comparative experiment based on even division and pose estimation is conducted in Table 4.

No matter what the value of even division number is, their performance is inferior to the uneven division. Compared with evenly dividing image into three pieces, 1.2% mAP gain can be obtained by the proposed method. If only using pose estimation to extract the local features, it can achieve 88.5% mAP and 78.0% mAP, which is lower than our method.

To further prove that our method can solve the problems of horizontal even division and pose estimation. This work compares our proposed method with other methods based on horizontal even division and pose estimation. The methods based on horizontal even division include HPM [54], MMHPN [55], LHF [56], and RANGE [57]. The methods based on pose estimation include DPA [58], HORelID [59], MBRAN [60], and DSA [61]. The result is illustrated in Table 5.

Obviously, compared with the methods based on horizontal even division and the methods based on pose estimation, our proposed method can obtain the best performance. Compared with the best methods listed in the table, the method proposed can obtain 2% mAP gain with 0.2% Rank-1 gain on Market-1501, and 1.7% mAP gain with 2.1% Rank-1 gain on DukeMTMC. This proves that our method can solve the problems of horizontal even division and pose estimation.

**Table 4 Evaluation on uneven division**

| Method              | Market-1501 | DukeMTMC |
|---------------------|-------------|----------|
|                     | mAP | Rank-1 | mAP | Rank-1 |
| Even division (×2)  | 88.0 | 94.1   | 77.1 | 86.4   |
| Even division (×3)  | 88.4 | 94.5   | 77.7 | 87.1   |
| Even division (×6)  | 88.8 | 94.9   | 77.9 | 87.6   |
| Pose estimation (CPM) | 88.5 | 94.7   | 78.0 | 88.1   |
| Uneven division (Ours) | 89.6 | 95.9   | 79.9 | 89.4   |

**Table 5 Evaluation on horizontal even division and pose estimation**

| Method          | Market-1501 | DukeMTMC |
|-----------------|-------------|----------|
|                 | mAP | Rank-1 | mAP | Rank-1 |
| HPM [54]        | 82.7 | 94.2   | 74.3 | 86.6   |
| MMHPN [55]      | 83.4 | 94.6   | 75.1 | 87.8   |
| LHF [56]        | 86.1 | 94.9   | 77.6 | 87.1   |
| RANGE [57]      | 86.8 | 94.7   | 78.2 | 87.0   |
| DPA [58]        | 85.6 | 95.2   | 73.1 | 86.5   |
| HORelID[59]     | 86.9 | 94.5   | 78.1 | 87.3   |
| MBRAN [60]      | 87.1 | 95.3   | 76.9 | 87.1   |
| DSA [61]        | 87.6 | 95.7   | 74.3 | 86.2   |
| Uneven division (Ours) | 89.6 | 95.9   | 79.9 | 89.4   |
Table 6 Evaluation on kernel function

| Kernel function | Market-1501 | DukeMTMC |
|-----------------|------------|----------|
|                 | mAP | Rank-1  | mAP | Rank-1  |
| \( \phi(v_p, v_g) = 1 \) | 88.1 | 95.5 | 78.1 | 87.6 |
| \( \phi(v_p, v_g) = \text{ave}(v_p, v_g) \) | 83.8 | 93.0 | 70.4 | 84.0 |
| \( \phi(v_p, v_g) = \max(v_p, v_g) \) | 87.7 | 95.0 | 77.4 | 87.6 |
| \( \phi(v_p, v_g) = \min(v_p, v_g) \) | 27.1 | 22.7 | 15.2 | 13.4 |
| \( \phi(v_p, v_g) = v_p \) | 21.9 | 17.3 | 12.2 | 10.1 |
| \( \phi(v_p, v_g) = v_g \) | 89.6 | 95.9 | 79.9 | 89.4 |

**Evaluation on kernel function**

The weight decision module takes the channel validity pair as input. A kernel function is utilized for processing input to generate the channel weight. Then, channel information fusion network combines five different channel weights with the channel similarity to infer the similarity. To evaluate the influence of different kernel functions on the performance, an experiment comparing different kernel functions is conducted. The results are shown in Table 6.

As shown in Table 5, different kernel functions achieve different performance. The first kernel function fixes the weight of each channel to 1. Equal-weight channel information fusion is performed in this condition, resulting in 88.1% mAP with 95.5% Rank-1 on Market-1501 and 78.1% mAP with 87.6% Rank-1 on DukeMTMC. Compared to others, the last kernel function has the best performance. It ignores the channel validity from gallery image, and considers the channel validity from the probe image as the channel weight. These results show that when testing, the channel information from the probe image should be paid more attention, as it is more significant.

**Ablation evaluation on blend metric loss**

The proposed blend metric loss consists of the improved i-TriHard loss, softmax loss and center loss. To verify the effectiveness of our loss, an ablation experiment based on different losses is conducted. The results are shown in Fig. 9.

Compared to softmax loss, gains of 3.9% mAP and 2.3% Rank-1 can be obtained in Market-1501. Gains of 3.7% mAP and 3.6% Rank-1 can be obtained in DukeMTMC. BML achieves 0.3% mAP with 0.3% Rank-1 gains and 0.4% mAP with 0.6% Rank-1 gains in two datasets compared to not using center loss. BML achieves 3.5% mAP with 1.9% Rank-1 gains and 3.4% mAP with 3.1% Rank-1 gains compared to not using i-TriHard loss. This shows that the introduction of i-TriHard loss and center loss can optimize network efficiently. That also verifies the effectiveness of our proposed blend metric loss.

Unlike the traditional TriHard loss, the proposed i-TriHard loss introduces dynamic adjustment to the hardest sample distance and the distance margin based on the distribution of extra images. To demonstrate the effectiveness of the improved strategy, the result from comparative experiments based on i-TriHard loss are illustrated in Table 7. TriHard is the traditional loss, “DP” means dynamically adjust the hardest positive sample distance, “DN” means dynamically adjust the hardest negative sample distance, and “DM” means dynamically adjust the distance margin.

As shown in Table 7, traditional TriHard loss without dynamic adjustment can only obtain 86.9% mAP with 94.7% Rank-1 on Market-1501 and 77.0% mAP with 86.7% Rank-1 on DukeMTMC, which is lower than our proposed i-TriHard loss. After a dynamic penalty is embedded, the method can achieve 2% mAP and 2.6% mAP improvement. Dynamically adjusting the distance margin saw even further improvements.

![Fig. 9 Ablation evaluation on blend metric loss. a Market-1501. b DukeMTMC](image-url)
of 0.7% mAP with 0.4% Rank-1 gains on Market1501 and 0.3% mAP with 0.9% Rank-1 gains on DukeMTMC.

To prove the effectiveness of this proposed i-TriHard loss, we add a new experiment applying TriHard loss and other improved variants of TriHard to train our network. The result is shown in Table 8. The compared losses include TriHard loss [8], Quadruplet loss [9], MSML [10], ALHSM [62]. Softmax loss and center loss are also combined with these losses to optimize the network.

As illustrated in Table 8, the proposed i-TriHard loss in this work can achieve the best performance improvement. Besides, compared with other variants of TriHard loss in the table, our designed loss does not need introduce more complex calculation and it is simple to implement, which proves the effectiveness of i-TriHard loss.

### Acknowledgements

This work is supported by National Key Research and Development Program of China (No. 2020YFC2007200).

### Declarations

#### Conflict of interest

The authors declare no conflict of interest exists.

#### Open Access

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

### References

1. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of IEEE CVPR, Columbus, pp 152–159
2. Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings of IEEE CVPR, Las Vegas, pp 1249–1258
3. Cheng D, Gong Y, Zhou S, Wang J, Zheng N (2016) Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of IEEE CVPR, Las Vegas, pp 1335–1344
24. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification by pose invariant deep metric learning with improved triplet loss. IEEE Access 6:68089–68095

25. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multicamera tracking. In: Proceedings of IEEE ICCV, Amsterdam, pp 17–35

26. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-guided feature alignment for occluded person re-identification. In: Proceedings of IEEE ICCV, Seoul, pp 542–551

27. Zheng W, Gong S, Xiang T (2011) Person re-identification by probabilistic relative distance comparison. In: Proceedings of IEEE CVPR, Providence, pp 649–656

28. Andriluka M, Pishchulin L, Gehler P (2014) 2d human pose estimation: new benchmark and state of the art analysis. In: Proceedings of IEEE CVPR, Providence, pp 3686–3693

29. Deng J, Dong W, Socher R, Li L, Li K, Li F (2009) ImageNet: a large-scale hierarchical image database. In: Proceedings of IEEE CVPR, Florida, pp 248–255

30. Chen J, Qin J, Yan Y, Huang L, Shao L (2020). Deep local binary coding for person re-identification by delving into the details. In: Proceedings of ACM international conference on multimedia, Washington, pp 3034–3043

31. Ristani E, Tomasi C (2018) Features for multi-target multi-camera tracking and re-identification. In: Proceedings of IEEE CVPR, Salt Lake City, pp 6036–6046

32. Sun Y, Zheng L, Yang Y, Tian Q (2018) Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of IEEE ECCV, Munich, pp 501–518

33. Dai Z, Chen M, Zhu S, Tan P (2018) Batch feature erasing for person re-identification and beyond. arXiv preprint arXiv:1811.07130

34. Tay C, Roy S, Yap K (2019) Aanet: attribute attention network for person re-identifications. In: Proceedings of IEEE CVPR, Califor-nia, pp 7127–7136

35. Ainanm J, Qin K, Liu G, Luo G (2020) Enforcing affinity feature learning through self-attention for person re-identification. ACM Trans Multimed Comput Commun Appl 16:1–22

36. Hao L, Gu Y, Liao X, Lai S, Jiang W (2019) Bags of tricks and a strong baseline for deep person re-identification. In: Proceedings of IEEE CVPRW, California, pp 1487–1495

37. Jin X, Lan C, Zeng W, Chen Z, Zhang L (2020) Style normalization and restitution for generalizable person re-identification. In: Proceedings of IEEE CVPR, Washington, pp 3140–3149

38. Wang G, Yang S, Liu H, Wang Z, Yang Y, Wang S, Yu G, Zhou E, Sun J (2020) High-order information matters: learning relation and topology for occluded person re-identification. In: Proceedings of IEEE CVPR, Washington, pp 6448–6457

39. Li H, Wu G, Zheng W (2021) Combined depth space based architecture search for person re-identification. In: Proceedings of IEEE CVPR, pp 6729–6738

40. Chen J, Jiang X, Wang F, Zhang J, Zheng F, Sun X, Zheng W (2021) Learning 3d shape feature for texture-insensitive person re-identification. In: Proceedings of IEEE CVPR, pp 8146–8155

41. Li Y, He J, Zhang T, Liu X, Zhang Y, Wu F (2021) Diverse part discovery: occluded person re-identification with part-aware transformer. In: Proceedings of IEEE CVPR, pp 2989–2907

42. Nguyen BX, Nguyen BD, Do T, Tijputra E, Tran Q, Nguyen A (2021) Graph-based person signature for person re-identifications. In: Proceedings of IEEE CVPR, pp 3492–3501

43. Zhong Z, Zheng L, Kang G, Li S, Yang Y (2017) Random erasing data augmentation. arXiv preprint arXiv:1708.04896

44. He L, Liang J, Li H, Sun Z (2018). Deep spatial feature reconstruction for partial person re-identification: alignment-free approach. In: Proceedings of IEEE CVPR, Salt Lake City, pp 7073–7082

45. He L, Sun Z, Zhu Y, Wang Y (2018) Recognizing partial biometric patterns. arXiv preprint arXiv:1810.07309

46. Miao J, Wu Y, Liu P, Ding Y, Yang Y (2019) Pose-guided feature alignment for occluded person re-identification. In: Proceedings of IEEE ICCV, Seoul, pp 542–551
47. Miao J, Wu Y, Yang Y (2021) Identifying visible parts via pose estimation for occluded person re-identification. IEEE Trans Neural Netw Learn Syst 99:1–11
48. Jia M, Cheng X, Zhai Y, Lu S, Ma S, Tian Y, Zhang J (2021) Matching on sets: conquer occluded person re-identification without alignment. In: Proceedings of AAAI, pp 1673–1681
49. Kiran M, Praveen R, Nguyen–Meidine L, Belharbi S, Blais–Morin L, Granger E (2021) Holistic guidance for occluded person re-identification. arXiv preprint arXiv:2104.06524
50. Sun Y, Xu Q, Li Y, Zhang C, Li Y, Wang S, Sun J (2019) Perceive where to focus: learning visibility-aware part-level features for partial person reidentification. In: Proceedings of IEEE CVPR, California, pp 393–402
51. He L, Wang Y, Liu W, Liao X, Zhao H, Sun Z, Feng J (2019) Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In: Proceedings of IEEE CVPR, Seoul, pp 8449–8458
52. Zheng K, Lan C, Zeng W, Liu J, Zhang Z, Zha Z (2021) Pose-guided feature learning with knowledge distillation for occluded person re-identification. In: Proceedings of ACM international conference on multimedia, Chengdu, pp 4537–4545
53. Yang J, Zhang J, Yu F, Jiang X, Zhang M, Sun X, Chen Y, Zheng W (2021) Learning to know where to see: a visibility-aware approach for occluded person re-identification. In: Proceedings of ICCV, pp 11885–11894
54. Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z, Huang T (2019) Horizontal pyramid matching for person re-identification. In: Proceedings of AAAI, Honolulu, pp 8295–8302
55. Zhang Y, Liu S, Qi L, Coleman S, Kerr D, Shi W (2020) Multilevel and multi-scale horizontal pooling network for person re-identification. Multimed Tools Appl 79(39):28603–28619
56. Zhang H, Si T, Zhang Z, Zhang R, Ma H, Liu S (2020) Local heterogeneous features for person re-identification in harsh environments. IEEE Access 8:83685–83692
57. Wu G, Zhu X, Gong S (2022) Learning hybrid ranking representation for person re-identification. Pattern Recognit 121:108239
58. Guo J, Yuan Y, Huang L, Zhang C, Yao J, Han K (2019) Beyond human parts: dual part-aligned representations for person re-identification. In: Proceedings of ICCV, Seoul, pp 3642–3651
59. Wang P, Zhao Z, Su F, Zu X, Boulgouris N (2021) HOReID: deep high-order mapping enhances pose alignment for person re-identification. IEEE Trans Image Process 30:2908–2922
60. Fang H, Chen J, Tian Q (2020) Multi-branch body region alignment network for person re-identification. In: Proceedings of international conference on multimedia modeling, Tokyo, pp 341–352
61. Zhang Z, Lan C, Zeng W, Chen Z (2019) Densely semantically aligned person re-identification. In: Proceedings of IEEE CVPR, Long Beach, pp 667–676
62. Wang Y, Wang Z, Jiang M, Chen L, Shen T, Zhang W (2020) Joint deep learning of angular loss and hard sample mining for person re-identification. J Intell Fuzzy Syst 38(1):417–426

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.