Fine-Grained Spatial Alignment Model for Person Re-Identification With Focal Triplet Loss

Qinqin Zhou, Bineng Zhong, Xiangyuan Lan, *Member, IEEE*, Gan Sun, Yulun Zhang, *Graduate Student Member, IEEE*, Baochang Zhang, *Member, IEEE*, and Rongrong Ji, *Senior Member, IEEE*

***Abstract*— Recent advances of person re-identification have well advocated the usage of human body cues to boost performance. However, most existing methods still retain on exploiting a relatively coarse-grained local information. Such information may include redundant backgrounds that are sensitive to the apparently similar persons when facing challenging scenarios like complex poses, inaccurate detection, occlusion and misalignment. In this paper we propose a novel Fine-Grained Spatial Alignment Model (FGSAM) to mine fine-grained local information to handle the aforementioned challenge effectively. In particular, we first design a pose resolve net with channel parse blocks (CPB) to extract pose information in pixel-level. This network allows the proposed model to be robust to complex pose variations while suppressing the redundant backgrounds caused by inaccurate detection and occlusion. Given the extracted pose information, a locally reinforced alignment mode is further proposed to address the misalignment problem between different local parts by considering different local parts along with attribute information in a fine-grained way. Finally, a focal triplet loss is designed to effectively train the entire model, which imposes a constraint on the intra-class and an adaptively weight adjustment mechanism to handle the hard sample problem. Extensive evaluations and analysis on Market1501, DukeMTMC-reid and PETA datasets demonstrate the effectiveness of FGSAM in coping with the problems of misalignment, occlusion and complex poses.**

***Index Terms*— Person re-identification, spatial alignment, focal triplet loss.**

Manuscript received October 8, 2019; revised May 8, 2020; accepted June 16, 2020. Date of publication June 29, 2020; date of current version July 13, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant U1705262, Grant 61972167, and Grant 61802135, in part by the National Key Research and Development Program under Grant 2017YFC0113000 and Grant 2016YFB1001503, in part by the Fundamental Research Funds for the Central Universities under Grant 30918014108, and in part by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) under Grant 202000012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jianbing Shen. *(Corresponding authors: Bineng Zhong; Rongrong Ji.)*

Qinqin Zhou and Bineng Zhong are with the Department of Computer Science and Technology, Huaqiao University, Xiamen 361021, China (e-mail: bnzhong@hqu.edu.cn).

Xiangyuan Lan is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong.

Gan Sun is with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China.

Yulun Zhang is with the Department of ECE, Northeastern University, Boston, MA 02115 USA.

Baochang Zhang is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China.

Rongrong Ji is with the Media Analytics and Computing Laboratory, Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China (e-mail: rrji@xmu.edu.cn).

Digital Object Identifier 10.1109/TIP.2020.3004267

I. Introduction

PERSON re-identification (re-ID) refers to retrieving a given probe image of pedestrian among a large set of gallery images captured by cameras with non-overlapping viewing angles. It has been a research hot spot in recent years, driven by the building of large-scale datasets [1]–[3] and remarkable baselines [4]–[7]. Despite the exciting progress, person re-ID retains as an open problem mainly due to the challenges like inaccurate detection, pose variation, background clutter, illumination change and occlusion.

In particular, complex poses, inaccurate detection, occlusion and misalignment are among the core challenges of person re-ID. To handle these issues, various robust features and models are proposed in recent works [4], [8]–[11], ranging from hand-crafted features, deep global features to local part based features. Based on the hand-crafted features, several traditional person re-ID methods [8], [12], [13] have made a certain progress. In addition, there were many methods [4], [9], [14] extracted global features from the entire image to address the issues of person re-ID. However, these methods often suffer from the issues of redundant backgrounds and misaligned body parts due to the lack of local cues. Recently, local partition methods [10], [11], [15], [16] have been introduced in person re-ID with remarkable performance gains. However, these local partition based person re-ID methods only utilize the coarse-grained level partitions, *e.g.* patches [15] or horizontal stripes [17], which often introduce redundant backgrounds and noises interference into the re-ID process. In addition to the methods of the above types, loss functions based methods [5], [18] deal with person re-ID problem from different perspectives, and put more attention on the constraints of distances among different samples. Although these methods have achieved competitive accuracy, they are still lack of consideration for the issues of hard sample problem in person re-ID and the constraint for absolute intra-class distance. For example, triplet loss defines the relative distance between anchor and inter-class, which uses the same weight to deal with each triplet. Therefore, the performance of these methods is often impaired by hard samples. Different from the above methods, in this work, we consider a spacial alignment on a fine-grained level and design a focal triplet loss to further improve the performance of person re-ID.

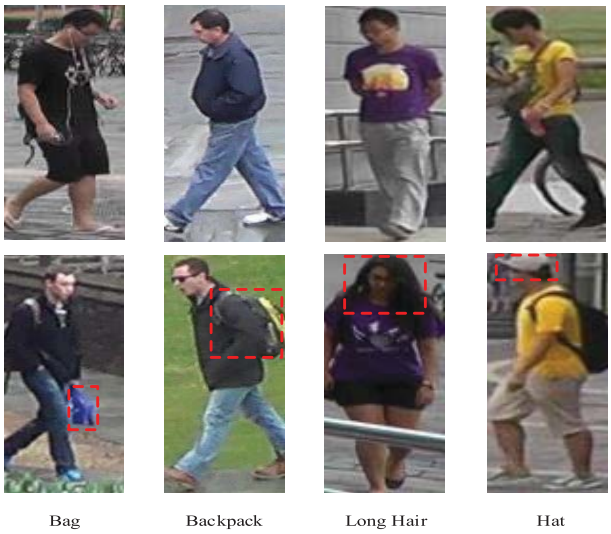


Fig. 1. The illustration of attributes in person re-ID, where each column shows two different people in a similar appearance, and each person can be easily distinguished by different attributes (*e.g.*, carrying bag, carrying backpack, the length of hair, and wearing hat).

In this paper, we propose a novel model named Fine-Grained Spatial Alignment Model (FGSAM) to tackle the above issues in a unified and explicit framework. Our framework jointly learns attribute and pixel-level local parts based on the pose information, which well discovers and embeds local context into deep global features towards robust person re-ID. In particular, FGSAM adopts attribute information with fine-grained local parts to improve the local perception of the model, as well as refining the alignment between different local parts. First, As illustrated in Fig. 1, human attribute helps to distinguish identities with similar appearances. Therefore, we use the attributes in the existing attribute-labeled datasets [1]–[3] in our FGSAM. Furthermore, a pose resolve net is proposed to encode pixel-level portrait spatial information with the channel parse block (CPB) to supervise the subsequent training of our FGSRM. This component enables our FGSRM to learn more useful fine-grained foreground information and reduce the impact of redundant backgrounds. Second, to tackle the misalignment problem, we propose a locally reinforced alignment mode, which discriminately combines different local parts with the corresponding attributes in different branches to align these local parts. As such multiple branches include different local parts and their accessory attributes, multiple embeddings are then used to represent one person. To better parse such a embedding space, we further propose a simple yet effective loss function termed focal triplet loss (FTL) to apply spatial constraints on intra-class and inter-class distances, and give larger weights to learn the hard samples. Notably, the proposed loss function gives an additional absolute distance constraint for intra-class and includes an adaptive weight adjustment mechanism to strengthen the learning of hard samples. Finally, a fine-grained level embedding space for intra-class and inter-class is learned in an end-to-end manner. To sum up, our contributions are three-fold:

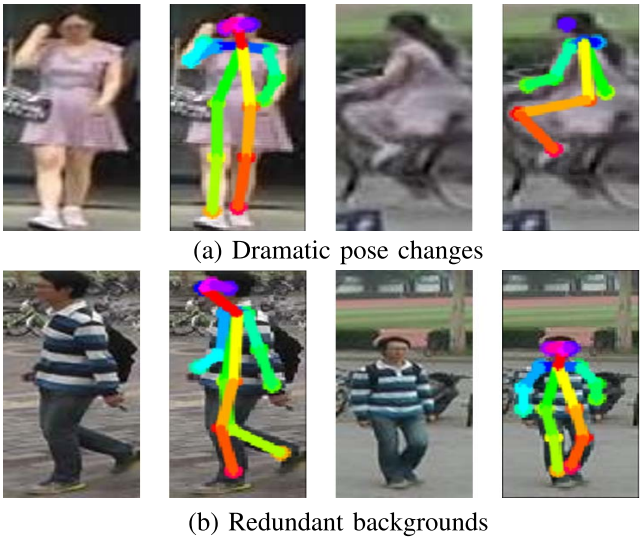
- We propose to optimize both intra-class and inter-class distances in a fine-grained way, which includes different body parts resolved from complex poses and their accessory attributes. To the best of our knowledge, this is the first attempt to explicitly exploits the pose and attribute information in an alignment way.
- We propose a FGSAM to embed pixel-level local parts and attribute information of human in a consistent representation. FGSAM includes a pose resolve net to obtain highlighted body features, and aligns the local parts with the proposed locally reinforced alignment mode.
- We design a novel focal triplet loss to optimize the absolute distances of intra-class, while adaptively increasing the weight for learning hard samples. Consequently, our model can be effectively trained regardless of the sample distribution imbalance among different classes in most person re-ID datasets.

This work is an extension of our conference version [19], with three new innovations: (1) We further investigate the alignment among fine-grained local parts, with a locally reinforced alignment mode to enhance the model robustness against the misalignment problem. (2) We propose a focal triplet loss to improve the model adaptability for the hard samples. (3) We add more comprehensive experiments to validate the locally reinforcement model and focal triplet loss, as well as an additional experiment on PETA [3], which includes several subsets collected from multiple different datasets.

The rest of this paper is organized as follows: Section II briefly presents the overview of the related work. Section III introduces our proposed FGSAM in details. The experimental results and corresponding analysis are detailed in Section IV. Finally, we conclude this paper in Section V.

II. Related Work

Person re-ID has received great attention in computer vision and beyond. As deep learning has been widely studied in many computer vision tasks [20]–[25], there are also many excellent works [4], [10], [11], [14], [15], [26]–[37] flooding into the research community of person re-ID. Most of these works are supervised learning, Ye *et al.* [36] proposed a Dynamic Graph Matching (DGM) framework to improve the label estimation process in an unsupervised manner. For the images captured by thermal and visible cameras, Ye *et al.* [37] proposed a novel bi-directional dual-constrained top-ranking (BDTR) loss to guide a dual-path network to learn discriminative feature representations. In [34], Zheng *et al.* proposed a joint learning framework to use the data generation technique to guide the re-id learning in an end to end way. The generated data may miss some local regions compared to the original data. In this paper, we focus on the supervised learning to build a re-ID system for the the images captured by ordinary cameras, and under the supervision of different labels to learn how to retrieve the pedestrian. In the past literature, global deep features based methods [4], [14] are widely used in person re-ID. Due to the human body is non-rigid, it is easy to cause misalignment for global feature based methods when human poses change seriously. Recent works in [11], [27] mainly focus on mining and integrating local features to include more discriminative information, with attribute information to increase the feature richness. Some works [38], [39] proposed the stereo matching techniques to search the target within a given area of



(a) Dramatic pose changes

(b) Redundant backgrounds

Fig. 2. Visualized examples of different poses in the person re-ID scenes, where each row shows different poses of one person. The original images are shown in the first and third columns. The same local part of one person in different images can be located with the resolved pose information.

the image, and the matching granularity is relatively coarse. In addition, there is another type of methods [18], [26] further committed to improving the loss functions to better construct the embedding space. In this section, we give a brief review of three most-related aspects in person re-ID, *i.e.*, coarse-grained level partition, attributes information and loss functions.

A. Coarse-Grained Level Partition

As the primary research object of person re-ID, human has the nature of non-rigid properties which causes misalignment and deformation problems. Based on this properties of human, there are many works [11], [16], [35], [40], [41] introducing coarse-grained level partition to cope with pose variations in person re-ID. As shown in Fig. 2, human pose changes and redundant backgrounds are common in person re-ID scenes and resolving the pose information can provide a better solutions for these situations. In [41], Zheng *et al.* proposed a PoseBox structure to obtain the features of different human body parts, and then fuses different grained of features to re-identify persons. Zhao *et al.* [16] designed a multi-branch deep neural network to capture the features of different human body parts, and fused these features to form the final feature representation. In [40], Kalayeh *et al.* introduced the human semantic parsing to obtain different local body parts and aggregated those parts in a consistent way to prevent misalignment. However, the partition strategies adopted by these methods are coarse-grained, which makes coarse positioning of the body parts that may lead to re-identification failures. In [35], Sun *et al.* proposed a general part-level feature learning method to describe an input with several part-level features, and a refined part pooling is designed to improve the original partition. Different from these methods, we propose a model which takes fine-grained local parts and attribute information of human in a consistent representation.

B. Attribute Information

Human attributes are widely available (such as in Market1501 [1], DukeMTMC-reid [2] and PETA [3]), which inspires us to integrate such cues in person re-ID. As shown in Fig. 1, when facing similar objects, one often distinguishes them by their distinguishable markers, which indicate that different attributes play an important role in person re-ID. In [42], Su *et al.* used a semi-supervised deep model to perform attribute learning based on global features. In [27], Feng *et al.* designed a deep network model with attribute learning to gain significant performance improvement on Market1501 and DukeMTMC-reid. In [4], Lin *et al.* used attribute information as a supplement to the identity recognition, which indicates that attribute information is beneficial to person re-ID. Considering both the global and local features are useful in person re-ID, a model simply adopts one of the global or local appearance feature is not enough to balance both sides perfectly. In this paper, we propose a FGSAM to integrate global feature with different fine-grained local feature, which interactively learns the attribute information. Our FGSAM includes two components to resolve global deep feature representation and complex human poses, respectively. Specifically, different local parts of a human body and their surroundings are the keys to identifying certain attributes, which indicates a spatial correlation between complex human poses and attributes. Based on this insight, we use different attributes to guide the learning of different fine-grained local parts. A channel parse block (CPB) is designed to learn to re-weight each channel of local features. Similar to the attention mechanism which has been widely studied in [43], [43]–[46], CPB aims at improving features' perception of important local parts. A locally reinforced alignment mode is further proposed to fuse global deep features of different persons with corresponding fine-grained local parts, which can boost the performance on pose resolving task.

C. Loss Functions

In addition to the improvement of person re-ID in the above two aspects, there are a certain amount of works [5], [18], [26], [47]–[49] which designed new loss functions to learn deep features for person re-ID. Traditional person re-ID methods use the cross-entropy loss to learn deep features of each identity, which ignores the metric information. To learn a better metric, Hermans *et al.* [26] proposed an online hard negative mining technique for the triplet loss. Triplet loss has been widely used in re-ID and other computer visual tasks like object tracking [50], [51]. As triplet loss focuses on the training data, its generalization capability is inadequate. To balance the generalization capability between the training set and testing set, Chen *et al.* [18] introduced a quadruplet loss. Compared to the triplet loss, such a quadruplet loss leads to a larger inter-class variation and a smaller intra-class variation. Different from the above methods, Wojke and Bewley [48] encoded the metric learning objective in a method for learning such a feature space, where the Cosine similarity is effectively optimized through a simple re-parametrization of

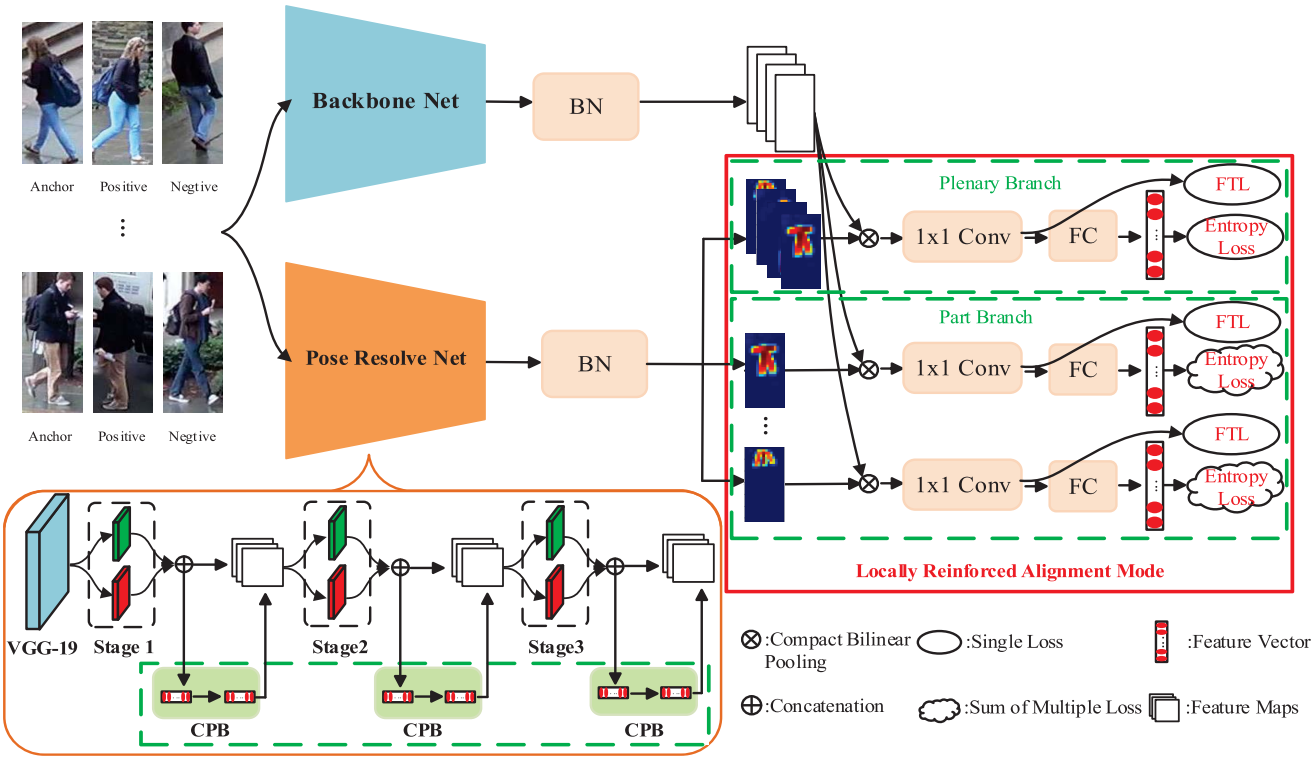


Fig. 3. The overview of our Fine-Grained Spatial Alignment Model, where "BN", "FC", "CPB" and "FTL" denote the batch normalization layer, full convolution layer, channel parse block and focal triplet loss, respectively.

the conventional softmax classification regime. These methods aim to construct a suitable embedding space for person re-ID, while ignoring the imbalance in data distribution that makes different sample learning costs being different. In this paper, we propose a focal triplet loss that takes both the absolute distance of intra-class and the imbalance in data distribution into consideration. In [49], Ye *et al.* proposed a weighted regularized triplet loss to retain the advantage of the triplet loss while avoiding the introduction of introducing any additional margin parameters. Our focal triplet loss is also based on the idea of weight, but we attach more importance to automatically balance the learning costs of different samples.

III. Fine-Grained Spatial Alignment Model

In this section, we firstly give an overview of our FGSAM. Then, we describe each component of our FGSAM in details. The motivation of our FGSAM focuses on fusing pose and attribute information of human to align the complex spatial distribution of a human body in a fine-grained way. Based on this idea, we firstly propose a pose resolve net which involves the proposed CPB to locate different parts of a human body at pixel level. Furthermore, we add attribute information to interactively learn the similarity of the aligned parts. The structure of our FGSAM is shown in Fig. 3. There are two main components in the chief part of FGSAM: a ResNet-50 as a backbone net and a pose resolve net to extract global semantic cues and pixel level parts cues of human, respectively. Based on our motivation, our FGSAM includes two complementary tasks: 1) identity recognition which includes one cross entropy loss and multiple focal triplet losses, 2) attribute discerning which includes multiple cross entropy losses. Given an image I as the input of the backbone net and the pose resolve net, the outputs can be respectively expressed as:

$$\mathcal{F} = \mathbb{M}(I), \tag{1}$$

Correspondingly, based on [52], the pose resolve net learns the part confidence maps and part affinity fields of local area on a human body which are produced by the final stage.

$$(\mathcal{H}^3, \mathcal{C}^3) = \mathbb{P}(I). \tag{2}$$

where $\mathcal{H}^3$ and $\mathcal{C}^3$ denote the part confidence map and the part affinity field obtained in the final stage. Based on the backbone net and the pose resolve net, we propose an elegant model for person re-ID and the detailed illustrations of these two main parts are given in following subsections. Moreover, the proposed focal triplet loss and a strategy combining multiple losses are presented.

A. Pose Resolve Net

The complex poses of human make person re-ID a tough computer vision task. To mitigate the impact of the complex poses, we propose a pose resolve net to discriminately resolve different local parts of a human body in an end to end way. In contrast to other pose based person re-ID methods, our pose resolve net considers the cases of occlusion and inaccurate detection. Specifically, we design an elegant module called channel parse block (CPB) to give different weights for different channels which corresponding to different local parts. To better explanation our pose resolve net, we introduce the backbone of our pose resolve net and CPB, respectively.

1) Backbone of Pose Resolve Net: Inspired by the success of the recently proposed pose estimation algorithm [52], we design a four-stage model as the backbone of our pose resolve net. The structure of our pose resolve net is shown in Fig. 3. Based on the good nature of bottom-up pose estimation model, pixel-level local parts can preliminary be located.

In our framework, the task of the pose resolve net is to capture the key parts of a human body in a fine-grained way and learn a channel-based parsing rule. More specifically, the two streams S_1 and S_2 in the pose resolve net provide part confidence maps and part affinity fields of the key parts, separately. In this work, we adopt the nineteen joints for a human body which will produce nineteen part confidence maps and thirty-eight part affinity fields. The pose branch is composed of four stages which include one VGG and three designed streams pre-trained on MSCOCO dataset [53]. Given an input image I, we denote the output part confidence maps and part affinity fields of the k-th stage as $\mathcal{H}^k$ and $\mathcal{C}^k$, respectively. It should be noted that the two streams in the pose resolve net are firstly trained with the MSCOCO dataset. More specifically, the pose resolve net is pre-trained with two mean squared error (MSE) loss functions for the two streams. Given the ground-truth maps $G_1^u(p)$, $G_2^v(p)$ and the estimated predictions $H_u^t(p)$, $C_v^t(p)$, the MSE losses for the two streams are formulated as follows:

$$\mathcal{L}_{S_1}^t = \sum_{u=1}^{U} \sum_{p} W(p) \cdot \left\| H_u^t(p) - G_1^u(p) \right\|_2^2, \tag{3}$$

$$\mathcal{L}_{S_2}^t = \sum_{v=1}^{V} \sum_{p} W(p) \cdot \left\| C_v^t(p) - G_2^v(p) \right\|_2^2. \tag{4}$$

To make use of pose information in pose resolve net, we aggregate the combination of the part confidence maps and part affinity fields of the pose resolve net with the feature maps of the backbone net by compact bilinear pooling. In this process, we add the batch normalization layer at the end of the backbone net and pose resolve net. The goal of this layer is to alleviate model overfitting and enforce the feature maps produced by these two nets to be subjected to the same distribution. Since the original structure in [52] is not designed for person re-ID, it is lack of consideration for the problems of occlusion and inaccurate detection. As the occlusion and inaccurate detection are common problems in person re-ID scenes, they cause the images of different identities usually include redundant backgrounds and the local parts of a human body usually become invisible. Therefore, how to suppress areas that are not related to the identities in the images and weaken the influence of the occluded part are another two key concerns of by our pose resolve net. Taking this as a starting point, we propose the CPB to handle the problems of occlusion and inaccurate detection. The details of the CPB are presented in the next.

2) Channel Parse Block: Different from previous coarse-grained level based person re-ID methods [10], [16], we propose a CPB which not only considers the fine-grained local parts, but also takes occlusion and inaccurate detection into account. Similar to the attention mechanism that widely studied in several works [43]–[45], which make use of the salient object annotation to directly train a deep model to predict the attentions on the targets. If the salient object annotation is available in the re-ID benchmarks, the attention on the pedestrian can be predicted, and the interference of complex background could further be reduced to some extent. Our CPB is to screen the channels of the features in the pose resolve net. For the lack of the salient object annotation, the CPB module re-weights each channel in the pose resolve net toward a goal that is conducive to retrieval in reid.

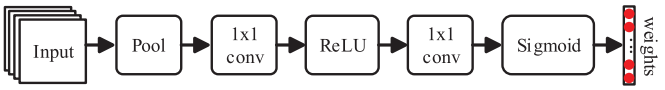


Fig. 4. The detailed structure of CPB.

Our CPB is inspired by the work of [54] from the fine-grained image recognition community which presents an architecture called SE-net block to weight different channels of features by modelling inter-dependencies between channels explicitly. Different from SE-net, our CPB is applied into the person re-ID community and we modify the FC layer in SE-net to 1×1 convolution layer to unify the pose resolve net and the CPB in the learning framework. The main goal of our CPB is to improve the pixel-level local parts outputted by the pose resolve net and make them more suitable with person re-ID. The detailed structure of the CPB is shown in Fig. 4. In our implementation, we use multiple CPBs to gradually learn the corresponding weights for different channels in the pose resolve net. As shown in Fig. 3, we aggregate the CPB with a multi-stage CNN to form our pose resolve net and train it to parse different channels of the feature maps outputted by the multi-stage CNN. The different channels of feature maps in our pose resolve net capture different local parts of a human body. Thus, our CPB not only learns to assign large weights to those channels in favor of identity recognition and attribute discerning, but also adaptively learns to suppress redundant backgrounds or occluded local parts.

B. Attribute Interactive Learning

In our FGSAM, we treat different attribute recognition tasks as multiple classification tasks. The ResNet-50 [55] is adopted as the backbone net of the other main part of our model. This is because ResNet-50 is suitable for the classification tasks. It should be noted that ResNet-50 can be easily replaced with other networks. Moreover, the pre-trained weights on the ImageNet are used for better convergence. As the pose resolve net focuses on the different local parts of a human body and their spacial relationships in person re-ID scenes, there lacks supervision information to guide the pose resolve net to learn effective local information during training phase. It is necessary to use another supervision mechanism that can be easily implemented in person re-ID scenes to make up for this defect. Therefore, we introduce attribute information which is available in several large person re-ID datasets (*e.g.*, Market1501, DukeMTMC-reid and PETA) to our FGSAM. For the attribute information is related with the local parts of

TABLE I

TABLE OF DIFFERENT ATTRIBUTES ADOPTED IN OUR FGSAM ON DIFFERENT DATASETS. L.SLV, L.LOW, L.UP, B.PACK, H.BAG, C.UP, C.LOW, C.SHOES DENOTE LENGTH OF SLEEVE, LENGTH OF LOWER-BODY CLOTHING, LENGTH OF UPPER-BODY CLOTHING, BACKPACK, HANDBAG, COLOR OF UPPER-BODY CLOTHING, COLOR OF LOWER-BODY CLOTHING, AND COLOR OF SHOES, RESPECTIVELY

Datasets	gender	hair	L.slv	L.low	L.up	T.low	hat	B.pack	bag	H.bag	age	C.up	C.low	C.shoes	boots
Market1501	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓		
Duke	✓				✓		✓	✓		✓	✓	✓	✓	✓	✓
PETA	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓

a human body, we use different attribute classification tasks to interactively learn the corresponding attentions on the local parts. That is the main idea of our attribute interactive learning strategy and we describe the details as follows.

Firstly, we use a compact bilinear pooling to fuse the outputs of the backbone net with the outputs of pose resolve net. As the compact bilinear pooling preserves the expression ability of the inputs, the rich semantic information outputted by the backbone net and pose resolve net can be effectively combined. Besides, the memory and computational consumptions of the compact bilinear pooling are lower than that of the ordinary matrix multiplication. Then, we adopt multiple cross entropy losses for the identity and attribute recognition tasks to train an intermediate model. Specifically, the attribute recognition tasks guide the pose resolve net to learn more discriminated local information and enhance the expression ability of the final fused features. In order to adapt to the re-ID tasks, we replace the FC layer at the end of the backbone net with N FC classification layers for attributes recognition and a FC layer for identity recognition. The attributes used in our framework on Market1501, DukeMTMC-reid and PETA are listed in Table I. It should be noted that PETA consists of multiple subdatasets and the labels of attribute up to 65. Therefore, we re-organize the labels of attributes in PETA to match the other two datasets. For the identity and the i-th attribute, we firstly use cross entropy loss to train the network as an intermediate model. Let $[[p^i(1), p^i(2), \cdots, p^i(N)]$ be the vector produced by the i-th softmax layer, we calculate the cross entropy loss of the i-th attribute and the identity as follow:

$$\mathcal{L}^i = -\sum_{n=1}^{N} log(p^i(n))q^i(n), \tag{5}$$

where $p(n) = \frac{exp(z_n)}{\sum_{i=1}^{N} exp(z_i)}$ denotes the predicted possibility of each attribute or ID label obtained by the softmax function. $q^i(n) = 1$, if the predicted result is same as the ground-truth.

As the outputs of the backbone net and pose resolve net have differences in magnitude, we normalize the outputs of two branches by using the batch normalization as follows:

$$\mathcal{F} = \frac{f^{(k)} - \mu_f}{\sqrt{\sigma_f^2 + \varepsilon}}, \tag{6}$$

$$\mathcal{H}^3 = \frac{h^{(k)} - \mu_h}{\sqrt{\sigma_h^2 + \varepsilon}}, \tag{7}$$

$$\mathcal{C}^3 = \frac{c^{(k)} - \mu_c}{\sqrt{\sigma_c^2 + \varepsilon}}. \tag{8}$$

where μ_* and σ_* denote the corresponding mean and variance of a mini batch, and ε is a parameter to prevent the denominator from 0. After normalization, these three feature maps can be constrained in the same magnitude.

After training the intermediate model, we use a locally reinforced alignment mode to reconstruct the model. Specifically, we use the features after the pool5 layer of the backbone net to fuse different channels of the features outputted by the pose resolve net. The trained weights of the intermediate model is used to initialize the corresponding layers of subsequent model. Different from the intermediate model, we introduce a locally reinforced alignment mode and the focal triplet loss to train our FGSAM which learns local parts alignment in a fine-grain style. Details about the locally reinforced alignment mode and focal triplet loss will be introduced in following subsections.

C. Locally Reinforced Alignment Mode

By effectively combing pose and attribute information, our FGSAM can learn a fine-grained representation of a human body which can improve the person re-ID performances. However, the spatial distribution of different local parts of a human body shows strong irregularity which may results in the misalignment problem. Directly fusing the highlighted human body may cause inconsistency between the fine-grained local parts of a human body. This inspires us to learn a locally reinforced alignment mode by considering the independence between different local parts of a human body. As we mentioned earlier, the different channels of features outputted by the pose resolve net depict different local parts of a human body. In this section, based on different channels corresponding to different local parts, we propose a locally reinforced alignment mode to learn the alignment between different fine-grained local parts.

Since we use CPBs to re-weight each channel in the pose resolve net, the pose resolve net outputs the visible and important channels corresponding to the parts of a human body automatically. Different from partition different local parts in a coarse-grained way, we take a locally reinforced alignment mode to address the misalignment problem among different local parts.

In our locally reinforced alignment mode, we distinguishably fuse different channels of the pose resolve net with the feature outputted by the backbone net. More specifically, we divide the channels of the pose resolve net into six parts (*i.e.*, head, upper body, upper arm, lower arm, upper leg and lower leg). Based on these parts, we assign different attributes to their related parts as a local supervision mechanism. In addition, we remain the branch which includes all parts for identity

recognition. For the different parts, there is a 1×1 convolution layer which achieves cross-channel information integration. As shown in Fig. 3, there are several branches in our locally reinforced alignment mode. Denote j-th local part of i-th human as $\mathcal{P}_i^j$, we adopt the compact bilinear to fuse each part with the outputs of the backbone net as six part branches. After the compact bilinear, there are six parallel 1×1 convolution layers which output the 256-dimensional embeddings after. In addition to the six part branches, we use all parts follow the same settings above to form the plenary branch which outputs a 512-dimensional embedding. These embeddings are used in the focal triplet loss for identity recognition.

Furthermore, we add a full convolution layer followed by a softmax layer to generate different predictions of attributes for the corresponding part branches and identity for the plenary branch. And the cross entropy loss is adopted as the identity loss for the plenary branch and attribute loss for different part branches.

As an important part in our locally reinforced alignment mode, the focal triplet loss is proposed to make up the deficiencies of the loss functions often used in person re-ID. On the one hand, the embedding space constructed by the triplet loss lacks constraints on absolute distance. On the other hand, the quantity of different identities shows serious imbalance which may cause the hard sample problem in the triplet loss. By addressing these problems, our focal triplet loss constructs a suitable embedding space for the identity recognition task. The detailed introduction of our focal triplet loss is presented in the next subsection.

As shown in Fig. 6, the different local parts defined in the locally reinforced alignment mode obtain corresponding weights in a fine-grained way. Specifically, the redundant backgrounds are suppressed effectively and weights are mainly concentrated on the local parts and attribute-related areas.

D. Focal Triplet Loss

In this section, we firstly introduce the original triplet loss and discuss the deficiency of triplet loss. Then, we describe the imbalance in the data distribution of most person re-ID datasets. Finally, we make a detailed description of our focal triplet loss to make up for these deficiencies.

The original triplet loss is proposed in [56] for the face recognition community and then introduced into the person re-ID. The triplet means the training data is divided into anchor, positive and negative to form a triplet structure. The relative distance between each element constitute the main parts of the loss function to optimize the models. Given a triplet as $\left(x_i^a, x_i^p, x_i^n\right)$ where x_i^a and x_i^p belong to the same identity, the triplet loss is defined as:

$$\mathcal{L}_{TL} = max\left(d\left(x_i^a, x_i^p\right) - d\left(x_i^a, x_i^n\right) + \alpha, 0\right), \tag{9}$$

where $d\left(x_i^a, x_i^p\right)$ and $d\left(x_i^a, x_i^n\right)$ denote the embedding distances of anchor from positive and negative, respectively.

The triplet loss gives a constraint for the relative distance between different samples to some extent. However, it ignores the absolute distance between different samples and the imbalance in data distribution which may learns an incorrect relative distance. As mentioned in [17], there often emerges large differences in the different intra-distances which may larger than inter-distances.

The imbalance in the data distribution exists in the most of person re-ID datasets which means the different identities of most datasets emerge large difference in quantity. As a result, the hard sample problem is common in person re-ID. There are several works have explored this problem. For example, Hermans *et al.* [26] proposed an online hard negative mining strategy that only considers binary weights for the most difficult sample pair in a batch. As a result, it may causes normal samples be susceptible to the outliers.

Based on the above analysis, we propose a focal triplet loss and a multi-loss joint strategy to guide our FGSAM to handle the hard sample problem. Our insight is to take the absolute distance of intra-class and adaptive adjustment weighting mechanism into consideration to design a proper loss function for our FGSAM. Thus, we add a constraint on the absolute distance of intra-classes and use the distance of intra-classes and inter-classes as a factor to adaptively adjust the weights for learning different samples. Specifically, we seek an adaptive way to give large weights for the hard samples. Inspired by the focal loss [57], we define $\lambda = max\left(d\left(x_i^a, x_i^p\right) - d\left(x_i^a, x_i^n\right) + 1, 1\right)$ as the indicator for hard and easy samples. Specifically, we define easy samples as $\lambda = 1$, otherwise we define difficult samples as $\lambda > 1$. Based on this indicator, the focal triplet loss is defined as follow:

$$\mathcal{L}_{FTL} = \lambda^{\gamma}\mathcal{L}_{TL} + max\left(d\left(x_i^a, x_i^p\right) - \beta, 0\right), \tag{10}$$

where λ^{γ} indicates the weight of the corresponding loss. When the sample is easy to distinguish from anchor, it means $\lambda < 1$ and the weight of the corresponding loss will be much smaller. Otherwise, the weight of the corresponding loss will be much larger when $\lambda > 1$. Besides, we use $max\left(d\left(x_i^a, x_i^p\right) + \beta, 0\right)$ as the absolute constraint within the class.

As shown in Fig. 3, our FGSAM uses multiple focal triplet losses and cross entropy losses for identity recognition and attribute recognition to jointly train the final model. Based on the multiple losses, we denote i-th attribute loss as $\mathcal{L}_{CEL}^i$ where $i = 1, 2, \cdots$ indicates the number of attributes in the corresponding datasets. The identity loss corresponding to the j-th branch is denoted as $\mathcal{L}_{FTL}^j$ where $j = 1, 2, \cdots$ indicates the number of branches in our FGSAM. In addition, there is a cross entropy loss for identity recognition mentioned in locally reinforced alignment mode which denoted as $\mathcal{L}_{CEL}^{ID}$, Therefore, the total loss of our FGSAM is computed as:

$$\mathcal{L} = \sum_i \mathcal{L}_{CEL}^i + \sum_j \mathcal{L}_{FTL}^j + \mathcal{L}_{CEL}^{ID}. \tag{11}$$

Since our FGSAM includes multiple losses, we use several phases to complete the training process for a better convergence.

IV. Experiments

In this section, we firstly introduce the datasets and the evaluation protocol adopted in our experiments. Then, we give a detailed description for our implementation. Thirdly,

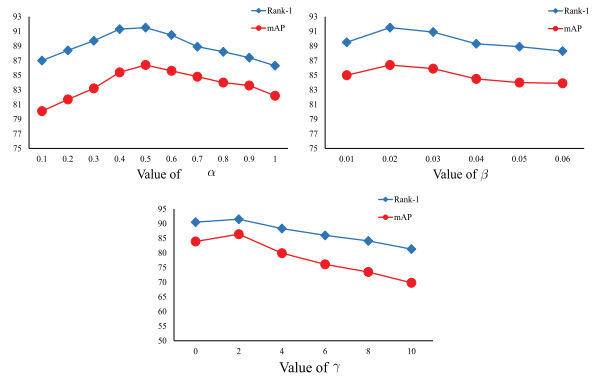


Fig. 5. The sensitivity analysis results of the hyperparameter α, β, γ in our focal triplet loss on Market1501 [1].

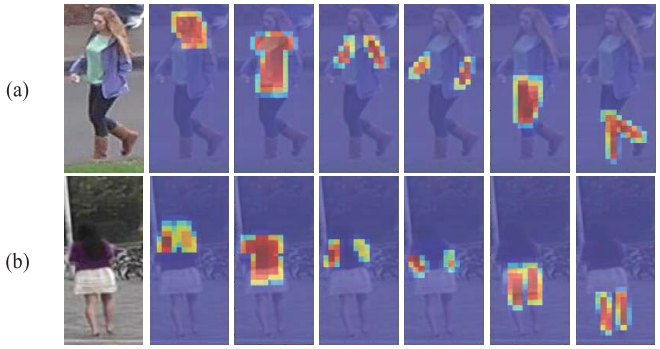


Fig. 6. The visualization results of different weights corresponding to different local parts which are defined in the locally reinforced alignment mode. (a) and (b) are two samples and the first column is the original image, the following six columns of images represent the different feature maps of the different local parts obtained from the pose resolve net.

we present the experimental results of our method on Market1501 [1], DukeMTMC-reid [2] and PETA [3] datasets when comparing with the state-of-art methods. Finally, we present the ablation analysis to verify the effectiveness of the different components in our method and the sensitivity of hyperparameters.

A. Datasets and Evaluation Protocol

1) Market1501 [1]: This dataset includes 1,501 identities captured by 6 different cameras. More specifically, there are 12,936 training images, 19,732 testing images and 3,368 query images detected by the DPM [66]. These images are with the resolution of 128×64, among which the training images and the testing images contain 751 and 750 identities, respectively.

2) DukeMTMC-reid [2]: DukeMTMC-reid is another large dataset which is used in multi-target tracking and person re-ID. It contains 16,522 training images, 17,661 testing images and 2,228 query images which cover 1,404 identities. All these images are captured at the Duke University campus by 8 synchronized cameras, then annotated and aligned manually.

3) PETA [3]: PETA is a large dataset for person attributes recognition and person re-ID. There are 19,000 images of different resolution and corresponding 61 binary attributes and 4 multi-class attributes collected from several subsets of other datasets. PETA is originally applied primarily to attribute recognition and needs a small amount of changes to accommodate person re-ID. In this paper, we follow the experimental protocol of ARP [4] to randomly select the same amount of training, query and test images of PETA.

4) Evaluation Protocol: In our experiments, we adopt the common evaluation protocol in person re-ID which includes the cumulative matching characteristic (CMC) curve and the mean average precision (mAP) for Market1501 and DukeMTMC-reid. The CMC curve is a precision curve that provides recognition precision for each rank. As a supplement, the mAP is the metric to measure the accuracy of person re-ID and it is the average of the maximum precisions at different recall values.

B. Implementation Detail

We use tensorflow to implement our FGSAM. The configuration of our computer is a HP Pro 3330 PC with Intel 1.8GHz×32 CPU and GTX-1080 GPU. As mentioned in Section III, we choose ResNet-50 as the baseline and we replace the FC layer after the pool5 layer with several 1×1 convolution layers for different classification tasks. We use a softmax layer to process the output of the 1×1 convolution layer to obtain the final classification results. For the pose resolve net, we use the MSCOCO dataset to re-train a four-stage model as the backbone.

In the training process, we divided it into several phases for a better convergence. In the first phase, we train the baseline with the initialized weights pre-trained on ImageNet. The baseline includes only a cross entropy loss for the identity recognition. The learning rate is initialized to 0.001 and decayed by 0.5 every 50 epoch. In this phase, the model is trained with 150 epochs in total. After the first phase, we use the compact bilinear pooling to combine the pose resolve net with the baseline. In this phase, we introduce multiple attributes recognition tasks into the model, and train this model with 60 epochs based on the first phase. More specifically, we add the batch normalization layer after the pool5 layer in baseline and the final stage of the pose resolve net to alleviate over-fitting. In this phase, we use the same learning rate and decay rate as the previous phase. 100 epochs are used to train our model. Finally, we use a locally reinforced alignment mode with multiple focal triplet losses to construct our FGSAM. Based on the number of attribute labels in different datasets, we adjust the number of branches in the locally reinforced alignment mode accordingly. We train this final model with 400 epochs. At this phase, we set the learning rate to 0.0001 and decay it by 0.5 every 100 epoch. In all training phases, the initialized weights of the latter phase are the trained weights corresponding to the previous phase. All training phases cost about 1.5 days and the testing phases cost about 1 hour. The Adam optimizer is implemented at the recommended parameters in each mini-batch to update the weights. The hyperparameters of α, β and γ in our focal triplet loss are set to 0.5, 0.01, 2, respectively. It should be noted that these phases in the training process are designed for better convergence of our FGSAM and there are no other fancy training tricks like random erasing augmentation in our training process.

TABLE II
QUANTITATIVE COMPARISON RESULTS ON MARKET1501 [1] AND DUKEMTMC-REID [2]

Methods	Market1501				DukeMTMC-reid			
	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP
Spindle [16]	76.9	91.5	94.6	-	-	-	-	-
PIE [41]	79.3	90.8	94.4	56.0	-	-	-	-
OIM [58]	82.1	-	-	-	68.1	-	-	-
ACRN [59]	83.6	92.6	95.3	62.6	72.6	84.8	88.9	52.0
SVDNet [60]	82.3	92.3	95.2	62.1	76.7	86.4	89.9	56.8
Part-aligned [61]	81.0	92.0	94.7	63.4	-	-	-	-
AACN [62]	85.9	-	-	66.9	76.8	-	-	59.3
ARP [4]	87.0	95.1	96.5	66.9	73.9	-	-	55.6
PSE [9]	87.7	94.5	96.8	69.0	79.8	89.7	92.2	62.0
GLAD [7]	89.9	-	-	73.9	-	-	-	-
MGCAM [6]	83.8	-	-	74.3	-	-	-	-
AANet-50 [63]	93.9	-	**98.6**	82.5	86.4	-	-	72.6
Lrdnn [19]	90.4	-	-	82.8	85.3	-	-	73.2
IANet [64]	**94.4**	-	-	83.1	87.1	-	-	73.4
MHN-6 (IDE) [65]	93.6	**97.7**	**98.6**	83.6	**87.5**	**93.8**	**95.6**	**75.2**
ours	91.5	96.8	97.3	**85.4**	85.9	90.2	93.4	74.1

TABLE III
QUANTITATIVE COMPARISON RESULTS ON PETA [3]

Methods	Rank-1	Rank-5	Rank-10	mAP
ARP [4]	58.1	78.3	75.7	55.8
Ours	**62.0**	**81.6**	**85.8**	**57.2**

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

Based on the evaluation protocol mentioned above, we compare FGSRM with the state-of-art methods on Market1501, DukeMTMC-reid and PETA, respectively. The comparison results on Market1501 and DukeMTMC-reid are shown in Table II. To the best of our knowledge, there are a small number of methods which belong to the similar category of ours and provide the comparable results on PETA. So we choose ARP [4] to compare with FGSRM and the details are shown in Table III.

As shown in Table II, we firstly adopt Spindle [16], PIE [41], OIM [58], ACRN [59], SVDNet [60], Part-aligned [61], AACN [62], ARP [4], PSE [9], GLAD [7], MGCAM [6], AANet-50 [63], Lrdnn [19], IANet [64] and MHN-6 (IDE) [65] to compare with our FGSRM on Market1501 and DukeMTMC-reid. It can be seen from this tables that our FGSRM achieves competitive results among the compared methods in rank-1 and mAP metrics. Specifically, our FGSRM exceeds MHN-6 (IDE) [65] by 1.8% in mAP on Market1501. For DukeMTMC-reid, FGSRM surpasses IANet [64] by 0.7% in mAP, and compared with MHN-6 (IDE) [65], FGSRM achieves comparable results.

As shown in Table III, our method outperforms ARP [4] with improvements up to 3.9% and 1.4% in rank-1 and mAP on PETA. Since PETA includes multiple person re-ID scenes, it indicates that our method has a certain adaptability in complicated scenes.

Among the compared methods, ARP [4] adopt attribute clues to guide the deep network to learn attribute recognition which helps re-identifying different persons and MGCAM [6] uses the pose clues to learn the local features of a human body. Different from these two methods, we aggregate both attribute and pose clues of a human body into one framework effectively. Moreover, we propose the locally reinforced alignment mode and the focal triplet loss to further enhance the performance of our FGSAM on person re-ID. As shown in Fig. 6, our method provides a fine-grained spatial alignment among different local parts of a human body. Overall, the comprehensive comparison results on the three datasets indicate the effectiveness of our method.

D. Ablation Analysis and Sensitivity Study of Parameters

To further verify the effectiveness of our approach, we compare the contributions of different components in our method by ablation analysis. For the pose resolve net, the attribute clues and the focal triplet loss, we perform an ablation study on both Market1501 and DukeMTMC-reid to verify their effectiveness. The ablation study of the CPB is performed on Market1501 for a clear comparison. Besides, there are several parameters in our focal triplet loss function and we present the sensitivity analysis in Fig. 5 on Market1501.

Firstly, we investigate the effect of the pose resolve net combined with the baseline and provide the results in Table IV. In this part, we add the batch normalization layer at the end of the pose resolve net and the baseline. Then, we adopt the compact bilinear pooling to obtain the fusion embeddings. As shown in Table IV, the pose resolve net (Base + Pose) obtains improvements in rank-1 and mAP on both Market1501 and DukeMTMC-reid. Furthermore, there are about 7.8% and 5.8% improvements in rank-1 and mAP on Market1501 compared to the baseline. This is consistent with our intuition that the pose resolve net helps to suppress redundant regions of images and effectively improves the representation ability of the embeddings.

Secondly, we use the locally reinforced alignment mode with attributes to find out how the locally reinforced alignment mode affects the accuracy. As shown in Table IV, there are promotions in rank-1 and mAP on both Market1501 and DukeMTMC-reid when the locally reinforced alignment mode (Base + Pose + LRAM) is adopted. Specifically, compared with the results of Base + Pose, there are about 7.9% and 1.4% improvements in mAP on Market1501 and DukeMTMC-reid, respectively. Based on the fusion features

TABLE IV

THE ABLATION ANALYSIS RESULTS OF DIFFERENT COMPONENTS IN OUR FGSAM ON MARKET1501 [1] AND DUKEMTMC-REID [2]. "BASE", "POSE", "LRAM" AND "FTL" MEANS BASELINE, POSE RESOLVE NET, LOCALLY REINFORCED ALIGNMENT MODE AND FOCAL TRIPLET LOSS, RESPECTIVELY

Methods	Market1501		DukeMTMC-reid	
	Rank-1	mAP	Rank-1	mAP
Base	80.3	69.8	74.1	65.5
Base + Pose	88.1	75.6	82.9	70.6
Base + Pose + LRAM	90.3	83.5	83.6	72.0
Base + Pose + LRAM + FTL	**91.5**	**86.4**	**85.9**	**74.1**

TABLE V

THE ABLATION ANALYSIS RESULTS OF THE CPB ON MARKET1501 [1]

Methods	Rank-1	mAP
Ours without CPB	89.9	85.4
Ours	**91.5**	**86.4**

from the combination of the pose resolve net and the baseline, the locally reinforced alignment mode brings attributes as supervisory information to further align different local parts in a fine-grained way which makes our FGSAM be robust to the complex poses and misalignment problems.

Thirdly, we analyze the effect of our focal triplet loss. As shown in Table IV, our focal triplet loss gets about 2.9% and 1.4% improvements in mAP on Market1501 and DukeMTMC-reid, respectively. There is also a certain improvement in rank-1. Compare with the original triplet loss and cross entropy loss, our focal triplet loss takes both absolute intra-class distance and hard sample problem into account which provides an adaptive weight adjustment mechanism for different samples and receives a more suitable embedding space.

Finally, we analyze the influence of the CPB on Market1501. As shown in Table V, based on the weight adjustment to the pose resolve net, we observe 1.6% and 1.0% improvements in rank-1 and mAP, respectively.

For the different parameters in our focal triplet loss function, the sensitivity studies of these parameters are shown in Fig. 5. Among these parameters, α controls the margin of intra-class and inter-class. β controls the absolute distance of inter-class. According to the results, our FGSAM achieves better results when $\alpha = 0.5$ and $\beta = 0.02$. It indicates that proper distance constraints for intra-class and inter-class help our FGSAM construct a suitable embedding space for preson re-ID. The exponential factor γ helps our FGSAM achieve better results when $\gamma = 2$. After the value of γ reaches a certain range, the performance begins to deteriorate. This mainly caused by the model biased in favor of learning hard samples when a excessive weight is given to the hard samples.

V. CONCLUSION

In this paper, we put our focus on the pose and attribute information of human and design a fine-grained spatial alignment model (FGSRM) for person re-ID with focal triplet loss. Different from existing person re-ID methods which separately combine pose or attribute information into person re-ID, we investigate how to jointly make use of pose and attribute information in person re-ID. For the complex poses of human, we introduce a pose resolve net to obtain local parts of a human body. As the inaccurate detection and occlusion problems make it difficult to obtain the pose information, we design a channel parse block in the pose resolve net, which can learn to further purify the pose information and suppress the redundant regions. For the attribute information, we analyze how to use the attribute information to assist the learning of pose information in a mutually reinforcing way, and further address the misalignment problem in person re-ID. Then, we propose a locally reinforced alignment mode which unites different attributes of human as a supervisory information to further align local parts in a fine-grained way. In addition, we investigate the deficiencies of the loss functions often used in person re-ID and propose the focal triplet loss. Our focal triplet loss makes up for the deficiencies of triplet loss and includes a mechanism of adaptive weight adjustment which learns to give greater weights for the hard samples. Finally, the experimental results on the Market1501, DukeMTMC-reid and PETA datasets have verified the effectiveness of our FGSAM. Since the image in the real scene may contain the identity of more than one pedestrian, the object detection algorithms are needed to extract these identities which require additional processing time. One possible direction for future research is use the stereo matching techniques to directly match the identities in different images without the help of the object detection algorithm, which can further improve the efficiency of reid.

REFERENCES

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

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

[3] Y. Deng, P. Luo, C. C. Loy, and X. Tang, "Pedestrian attribute recognition at far distance," in *Proc. ACM Int. Conf. Multimedia (MM)*. New York, NY, USA: ACM, 2014, pp. 789–792.

[4] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

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

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

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

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

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

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

[11] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3980–3989.

[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2360–2367.

[13] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 536–551.

[14] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2019.

[15] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3908–3916.

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

[17] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1335–1344.

[18] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, vol. 2, no. 8.

[19] Q. Zhou, B. Zhong, X. Lan, G. Sun, Y. Zhang, and M. Gou, "LRDNN: Local-refining based deep neural network for person re-identification with attribute discerning," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1–7.

[20] J. Shen, Z. Liang, J. Liu, H. Sun, L. Shao, and D. Tao, "Multiobject tracking by submodular optimization," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 1990–2001, Jun. 2019.

[21] J. Shen, D. Yu, L. Deng, and X. Dong, "Fast online tracking with detection refinement," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 162–173, Jan. 2018.

[22] X. Dong, J. Shen, W. Wang, L. Shao, H. Ling, and F. Porikli, "Dynamical hyperparameter optimization via deep reinforcement learning in tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Nov. 29, 2019, doi: 10.1109/TPAMI.2019.2956703.

[23] Z. Liang and J. Shen, "Local semantic Siamese networks for fast tracking," *IEEE Trans. Image Process.*, vol. 29, pp. 3351–3364, 2020.

[24] H. Hu, B. Ma, J. Shen, H. Sun, L. Shao, and F. Porikli, "Robust object tracking using manifold regularized convolutional neural networks," *IEEE Trans. Multimedia*, vol. 21, no. 2, pp. 510–521, Feb. 2019.

[25] Q. Zhou, B. Zhong, Y. Zhang, J. Li, and Y. Fu, "Deep alignment network based multi-person tracking with occlusion and motion reasoning," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1183–1194, May 2019.

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

[27] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.

[28] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 402–419.

[29] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2018, pp. 508–526.

[30] C. Wang, Q. Zhang, C. Huang, W. Liu, and X. Wang, "Mancs: A multi-task attentional network with curriculum sampling for person re-identification," in *Proc. ECCV*, 2018, vol. 2, no. 7, p. 8.

[31] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7073–7082.

[32] X. Zhang *et al.*, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: http://arxiv.org/abs/1711.08184

[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[34] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.

[35] Y. Sun, L. Zheng, Y. Li, Y. Yang, Q. Tian, and S. Wang, "Learning part-based convolutional features for person re-identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 5, 2019, doi: 10.1109/TPAMI.2019.2938523.

[36] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Dynamic graph co-matching for unsupervised video-based person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2976–2990, Jun. 2019.

[37] M. Ye, X. Lan, Z. Wang, and P. C. Yuen, "Bi-directional center-constrained top-ranking for visible thermal person re-identification," *IEEE Trans. Inf. Forensics Security*, vol. 15, no. 6, pp. 407–419, Jun. 2020.

[38] G.-Y. Nie *et al.*, "Multi-level context ultra-aggregation for stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3283–3291.

[39] M. Hallek, F. Smach, and M. Atri, "Real-time stereo matching on CUDA using Fourier descriptors and dynamic programming," *Comput. Vis. Media*, vol. 5, no. 1, pp. 59–71, Mar. 2019.

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

[41] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.

[42] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 475–491.

[43] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[44] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Mar. 18, 2019, doi: 10.1109/TPAMI.2019.2905607.

[45] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1531–1544, Jul. 2019.

[46] J. Shen, X. Tang, X. Dong, and L. Shao, "Visual object tracking by hierarchical attention Siamese network," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3068–3080, Jul. 2020.

[47] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 4321–4329.

[48] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 748–756.

[49] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193

[50] X. Dong and J. Shen, "Triplet loss in Siamese network for object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 459–474.

[51] X. Dong, J. Shen, D. Wu, K. Guo, X. Jin, and F. Porikli, "Quadruplet network with one-shot learning for fast visual object tracking," *IEEE Trans. Image Process.*, vol. 28, no. 7, pp. 3516–3527, Jul. 2019.

[52] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[53] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 740–755.

[54] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

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

[56] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.

[57] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[58] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3415–3424.

[59] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 20–28.

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

[61] L. Zhao, X. Li, Y. Zhuang, and J. Wang, "Deeply-learned part-aligned representations for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3219–3228.
[62] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.
[63] C.-P. Tay, S. Roy, and K.-H. Yap, "AANet: Attribute attention network for person re-identifications," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7134–7143.
[64] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-And-Aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.
[65] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.
[66] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

Qinqin Zhou received the B.S. degree in mathematics and applied mathematics from Qingdao Technological University, Shandong, China, in 2016, and the M.S. degree from the School of Computer Science and Technology, Huaqiao University, Xiamen, China, in 2019. His current research interests include pattern recognition, machine learning, and computer vision.

Bineng Zhong received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2004, 2006, and 2010, respectively. From 2007 to 2008, he was a Research Fellow with the Institute of Automation and the Institute of Computing Technology, Chinese Academy of Sciences. From September 2017 to September 2018, he was a Visiting Scholar with Northeastern University, Boston, MA, USA. He is currently a Professor with the School of Computer Science and Technology, Huaqiao University, Xiamen, China. His current research interests include pattern recognition, machine learning, and computer vision.

Xiangyuan Lan (Member, IEEE) received the B.Eng. degree in computer science and technology from the South China University of Technology, Guangzhou, China, in 2012, and the Ph.D. degree from the Department of Computer Science, Hong Kong Baptist University, Hong Kong, in 2016. He is currently a Research Assistant Professor with Hong Kong Baptist University. His current research interests include intelligent video surveillance and biometric security.

Gan Sun received the B.S. degree from Shandong Agricultural University in 2013 and the Ph.D. degree from the Shenyang Institute of Automation, Chinese Academy of Sciences, in 2020. He is currently an Assistant Professor with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences. He has been visiting Northeastern University from April 2018 to May 2019 and the Massachusetts Institute of Technology from June 2019 to November 2019. His current research interests include lifelong machine learning, multitask learning, medical data analysis, deep learning, and 3D computer vision.

Yulun Zhang (Graduate Student Member, IEEE) received the B.E. degree from the School of Electronic Engineering, Xidian University, China, in 2013, and the M.E. degree from the Department of Automation, Tsinghua University, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, Northeastern University, USA. His research interests include image restoration and deep learning. He was a recipient of the Best Student Paper Award at the IEEE International Conference on Visual Communication and Image Processing in 2015.

Baochang Zhang (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, 2001, and 2006, respectively. From 2006 to 2008, he was a Research Fellow with The Chinese University of Hong Kong, Hong Kong, and Griffith University, Brisbane, Australia. He is currently a Professor with the Science and Technology on Aircraft Control Laboratory, School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. He was supported by the Program for New Century Excellent Talents in the University of Ministry of Education of China. His current research interests include pattern recognition, machine learning, face recognition, and wavelets.

Rongrong Ji (Senior Member, IEEE) is currently a Professor and the Director of the Intelligent Multimedia Technology Laboratory and the Dean Assistant with the School of Information Science and Engineering, Xiamen University, Xiamen, China. His work mainly focuses on innovative technologies for multimedia signal processing, computer vision, and pattern recognition, with over 100 papers published in international journals and conferences. He is a member of ACM. He also serves as a program committee member for several Tier-1 international conferences. He was a recipient of the ACM Multimedia Best Paper Award and the Best Thesis Award of the Harbin Institute of Technology. He serves as an Associate/Guest Editor for international journals and magazines such as *Neurocomputing*, *Signal Processing*, *Multimedia Tools and Applications*, the *IEEE MultiMedia Magazine*, and *Multimedia Systems*.