Devil’s in the Detail: Graph-based Key-point Alignment and Embedding for Person Re-ID

Xinyang Jiang¹, Fufu Yu*, Yifei Gong¹, Shizhen Zhao³, Xiaowei Guo¹, Feiyue Huang¹, Wei-Shi Zheng², Xing Sun¹
¹ Tencent Youtu Lab, Shanghai, China
² Sun Yat-sen University, Guangzhou, China
³ Huazhong University of Science and Technology, Wuhan, China

{sevjiang, fufuyu, yifeigong, scorpioguo, garyhuang, winfredsun}@tencent.com
zhaosz@hust.edu.cn wszheng@ieee.org

Abstract

Although Person Re-Identification has made impressive progress, difficult cases like occlusion, change of view-point and similar clothing still bring great challenges. Besides overall visual features, matching and comparing detailed local information is also essential for tackling these challenges. This paper proposes two key recognition patterns to better utilize the local information of pedestrian images. From the spatial perspective, the model should be able to select and align key-points from the image pairs for comparison (i.e. key-points alignment). From the perspective of feature channels, the feature of a query image should be dynamically adjusted based on the gallery image it needs to match (i.e. conditional feature embedding). Most of the existing methods are unable to satisfy both key-point alignment and conditional feature embedding. By introducing novel techniques including correspondence attention module and discrepancy-based GCN, we propose an end-to-end ReID method that integrates both patterns into a unified framework, called Siamese-GCN. The experiments show that Siamese-GCN achieves state-of-the-art performance on three public datasets.

1. Introduction

Person re-identification (ReID) increasingly draws attention due to its wide applications in surveillance, tracking, smart retail, etc. Although ReID methods progress rapidly and achieve impressive performance on benchmark datasets, in practice, difficult cases like occlusion, change of view-point and similar clothing still bring great challenges. As shown in Figure 1a), in these cases, the overall appearance of a pedestrian may not always be reliable, and the detailed local feature becomes essential. Thus, this paper focuses on how to effectively utilize the detailed information for matching pedestrian images.

Looking at how human annotators would compare the similarities between two images, we find that there are two key recognition patterns involving matching with detailed local features. Firstly, when the overall visual similarity of the entire body is unreliable for matching a pair of images (e.g., Figure 1a)), a human annotator will select several local regions decisive for recognition and align the corresponding regions between two images for feature comparison. For example, in Figure 1b), key-points pairs including hat, shoulder, arms, and shoes are selected for comparison, and since all of these pairs have high feature similarity, the model will have high confidence to accept the images as the same person. The same logic can be applied to negative examples, as shown in Figure 1c), the general appearance of the image pair is very similar, but one can reject the images as negatives by comparing local feature similarity around the decisive key-point pairs including the head, legs, shoes and coat pockets.

Secondly, for the same query image, a human annotator’s attention to visual features varies drastically when matching with different gallery images. Figure 2 is an intuitive exam-
Figure 2. Conditional Feature Embedding: The illustration of changing conditional feature when comparing an image with different contextual images.

ple to explain this recognition pattern. For the same query image in Figure 2, different feature vectors need to be extracted to match gallery image A and gallery image B. Since the face and glasses cannot be seen in gallery image A, in the feature vector of the query image, the channels related to these semantics are suppressed. Similarly, when matching query image with gallery image B, channels related to the black jacket and plastic bag are suppressed.

In conclusion, a good ReID matching model should meet two requirements:

- Spatially, sub-regions decisive for recognition need to be selected and aligned for feature comparison. We call this key-point alignment.
- Channel-wise, the feature extraction of a query image should be dynamically adjusted based on the gallery image it matches. We call this conditional feature embedding.

Many existing state-of-the-art methods boost the ReID performance by utilizing local detailed information. Next we examine if these methods meet above-mentioned two requirements. Table 1 summarizes how these methods meet the requirements. Although local feature learning methods [21] [10] [28] [31] [18] significantly boost ReID performance by learning features of local body regions, they do not select and align learned local features but directly fuse all features into a unified vector. Some recent alignment-based methods propose to align the local parts between image pair based on human part visibility [19] or part similarity [20] [26] [5]. The overall similarity of the image pair is decided by the local similarity of the aligned part pairs. The local features of the alignment-based methods are pre-extracted from the individual images and is not adjustable to different contextual images. Secondly, we argue that using predefined alignment rules (e.g., local feature’s similarity or same human part) is not flexible enough to select suitable key-point pairs for different image pairs. As shown in Figure 1b, several key-point pairs with high visual similarity are selected, while in Figure 1c key-point pairs with large visual difference should be selected to reject the image pair. Instead of predefined rules, we propose to learn a novel correspondence attention module to automatically select decisive key-point pairs based on the visual content of both images. Recently Wu et al. proposes a series of ReID methods to learn conditional feature embedding (i.e. DCCs [24] and Deep Spatially Multiplicative Integration [25]). They learn conditional features by feeding a pair of fused global feature vectors into a RNN, but key-point alignment is not applied. Furthermore, we argue that RNN, although excels at modeling linear relation in a sequential structure, is less suitable to model the complex topological structure of local region correspondence.

In conclusion, as shown in Table 1 to the best of our knowledge, existing methods are not able to apply both conditional feature embedding and key-point alignment into the same model. Thus, we propose a novel graph-based ReID model that integrates both recognition patterns into a unified framework, called Siamese-GCN. As shown in the orange box of Figure 3, for key-point alignment, instead of predefined alignment rules, we propose a novel correspondence attention module to dynamically select and align key-points both within an image and between image pair. Secondly, as shown in grey box of Figure 3, our model takes the obtained key-point correspondence graph as inputs and extracts conditional feature embedding with a novel graph convolutional network [9] [1]. The reason we use GCN is because, the key-point correspondence obtained by our attention module forms very complex graph including many-to-many connections both inside an image and between two images, and GCN is able to capture the the high ordered topological relation among multiple key-points. However, the standard GCN smooth the adjacent node features, which does not meet the requirement of a recognition task to compare the feature difference between key-points. Hence, a novel discrepancy-based graph convolution is proposed.

The contribution of our proposed method is listed as follows:

- Siamese-GCN is able to integrate both key-point alignment and conditional feature embedding into a unified ReID framework.
- Instead of using a pre-defined Adjacency Matrix like standard GCN, our Siamese-GCN uses a novel correspondence attention module where the relation graph is automatically predicted and dynamically adjusted during training.
Table 1. The comparison of state-of-the-art ReID methods on meeting recognition patterns of key-point alignment and conditional feature embedding.

| Category          | Methods               | Key-point Align. | Cond. Feat. Embed. |
|-------------------|-----------------------|------------------|--------------------|
| Local Feat. Learn.| Stripe: MGN [21], Pyramid [30] | N/A              | N/A                |
|                   | Human Parsing: PIE [31]| N/A              | N/A                |
|                   | Attention: ABDNet [2]  | N/A              | N/A                |
| Alignment Based   | AlignedReID [13]      | Similarity-based | N/A                |
|                   | VPM PCB [20], [19]    | Part Visibility  | N/A                |
| Joint Feat. Learn.| DCCs [24]             | N/A              | RNN-based          |
|                   | Our method            | Corres. Att.     | discrepancy-based GCN |

- Compared to standard GCN that smooths the feature of adjacent nodes, we propose a novel graph convolution that meets the requirement of pairwise matching scenario for ReID and computes the discrepancy between adjacent graph nodes.

2. Related Works

2.1. Part base methods

Part-based models learn local features of different body parts to enhance the global ReID feature on cross-view matching. One of the most common and effective type of part-based models simply split the output feature-maps of ReID model’s intermediate layers into several horizontal stripes and learn local features for each stripe, such as PCB [20], MGN [21] and Pyramid [30]. VA-ReID [35] proposes a view-aware angular loss to train the part-based model, but it uses a much larger backbone network (ResNext101) compared to ours, so we do not include this method in our performance comparison. Another type of part-based models [10], [28], [31] segment human body into meaningful body parts and learn local feature for each body parts. SPRReID [8] learns a human parsing branch for body part segmentation and fuses local features for different parts by weighted average pooling. DSA-ReID [27] proposes projects human parts into a UV space and uses this UV space branch to guide the learning of a stripe model.

2.2. Alignment-based methods

On top of the local features, instead of fusing the local features directly, some methods propose to align parts from a pair of images, and match a pair of images based on the similarity of their aligned part pairs. AlignReID [26] proposes a dynamic programming algorithm to align a stripe in the image to a stripe in another image based on their local feature similarity. DSR [5] and SFR [6] propose a sparse coding method to implicitly look for similar keypoint pairs by reconstructing one image’s feature map with another. VPM [19] proposes to align the stripes from two images based on the visibility of each stripes. PGFA [14] exploits pose landmarks to align stripes. CDPM [22] proposes to localize and align local parts by a sliding window method. Some GAN based methods like FD-GAN [3] propose to align local features by directly transfer the image to the same pose and viewpoint of the target image.

2.3. Joint Feature Learning

In this paper, Joint Feature Learning refers to methods that feed both image into a model simultaneously to obtain a conditional feature embedding. Existing methods like DCCs [24] and Deep Spatially Multiplicative Integration [25] learns a RNN to iteratively encode couple features step-by-step.

2.4. Attention-based methods

Attention based methods are a type of State-of-the-Art ReID methods that propose to select important regions or channels of a feature-maps to form the ReID feature and discard region irrelevant to recognition such as background. Unlike our Siamese-GCN that selects decisive region pairs based on a pair of images, most of the existing attention based methods focus on selecting important information from individual images. Method in [29] proposes to predict multiple attention maps for different human parts. HA-CNN [13] uses a Harmonious Attention module to conduct feature selection both spatially and in channel-wise for a individual image. ABDNet [2] proposes a similar spatial and channel attention with an orthogonality constraint.

2.5. Graph-based methods

Recently some methods propose to use graph-based techniques to learn more complex relationship for ReID model. However, instead of exploring the complex correlation between detailed local regions inside image pairs, most of the existing methods are based on global features. SG-GNN [16] use graph to represent the relation between multiple probe images and gallery images and use a graph neural network update samples’ global features with a massage passing method. Group shuffling random walk [17] further extend the probe-gallery relation to gallery-gallery relationship.
3. Methods

3.1. General Framework

Figure 3 shows the general training workflow of Siamese-GCN. Siamese-GCN is an end-to-end learning framework containing three stages, namely individual feature embedding, decisive key-point alignment, and conditional feature embedding. At the first stage, our method extracts individual feature maps for both images. Secondly, given the feature-maps of two images, a correspondence attention module is used to select decisive key-point pairs both within an image and between image pair. Finally, a novel discrepancy-based GCN is used to extract conditional features from the key-point correspondence graph. The following three sub-sections elaborate on the implementation and formulation the three stages.

3.2. Individual Feature Embedding

The individual feature embedding stage is responsible for extracting a feature-map for each individual pedestrian image. Any type of CNN backbones for person Re-identification can be applied for the individual feature extraction.

As shown in the blue box in Figure 3 to enforce the backbone network extracting good individual features, an additional training loss branch is attached to the module. We use an feature encoder to encode the feature-map extracted by the backbone network into individual feature vectors $v$ by applying a Global Average Pooling followed by a $1 \times 1$ convolution layer. Then, a cross entropy based ID loss is used to train the individual feature extractor:

$$L_{CE}(y, v) = \frac{1}{C} \sum_{c} C_{y} \log(p(c|v))$$

(1)

where, $p(c|v) = \frac{e^{W_{c}v}}{\sum_{c} e^{W_{c}v}}$, and $W$ is a weight matrix of a fully connect layer to classify $v$ into different identities, and $C$ is the total number of identities in the training set.

3.3. Key-point Alignment

In order to decide if a pair of images has the same identity, several decisive region pairs are selected where their local features are compared. We call this process the decisive key-point alignment. Given a feature-map of an image extracted at individual feature learning stage denoted as $X \in R^{H \times W \times c}$, where $H$ and $W$ is the height and width of the feature-map and $D$ is the number of feature channels, the selected pairs of pixels in $X$ forms a undirected graph, whose adjacent matrix $A$ will be used to extract conditional features of both images with Graph Convolutional Network.

3.3.1 Predefined Alignment Rule

In our Siamese-GCN framework and many other alignment-based methods, the decisive key-point alignment is conducted on the feature-maps extracted by the previous Individual Feature Embedding stage. One of the common way to obtain the correlation between two feature-map pixels is to compute feature similarity. For example, given a pair of feature-maps each containing $HW$ pixels, a cosine similarity can be used to evaluate the likelihood that a pixel pair $x_i, x_j$ is correlated:

$$S_{ij} = d(x_i, x_j)$$

(2)

Following a similar strategy in [26], each pixel in one feature-map is assigned to a pixel with the highest similarity from the other feature-map, which forms a decisive key-point pair. As a result, the adjacent matrix $A_{ij}$ between any pixel from two feature-maps are formulated as:

$$A_{ij} = \begin{cases} 0 & j \neq \arg \max_k d(x_i, x_j) \\ S_{ij} & j = \arg \max_k d(x_i, x_j) \end{cases}$$

(3)

Noted that to provide more information on the correlation level of a key-point pair, we assign the edge of the key-point pair with a similarity weight, instead of a binary value.

Another way to build the pixel correspondence between feature-map pairs is by the human part or other semantic labels of each pixel. Given a pair of pixels indexed by $i$ and $j$ with human part label $a_i, a_j$, the adjacent matrix is formulated as:

$$A_{ij} = \begin{cases} 0 & a_i \neq a_j \\ 1 & a_i = a_j \end{cases}$$

(4)

3.3.2 Correspondence Attention Module

As discussed in Section 1, previous two types of methods for key-point alignment and selection are based on predefined rules and are not flexible enough for different types of image pairs. We propose to train a correspondence attention module to automatically select decisive key-point pair.

Our method not only focuses on finding decisive key-point pairs between two images, but also discovering intra-relationship between key-points within an individual image. Given feature-map the individual feature-map $X \in R^{H \times W \times C}$, we first reshape the feature-map into a $HW \times c$ matrix denoted as $X_r$. Then the correspondence of any pair of pixels in $X$ is computed as follows:

$$S = X_rW^T$$

(5)

where $W$ is a diagonal parameter matrix to assign a learnable weight for each dimension of feature-map channel.
Similarly, the inter-image correspondence of two different feature-maps $X$ and $X'$ is:

$$S' = X_w W' X_w^T$$

We combine both intra-image correspondence and inter-image correspondence to obtain the overall adjacent matrix of the key-point pair:

$$A = ReLU \left( \begin{array}{cc} S_i & S_i' \\ S_j & S_j' \end{array} \right)$$

where a ReLU activation is used to clip the value below zero to zero in the combined adjacent matrix to obtain a more sparse key-point pair with edges weighted by positive values.

### 3.4. Conditional Feature Embedding

Given a pair of images to be matched, denoted as $(I_i, I_j)$, our goal is to extract conditional feature embedding for one image conditioned on the other, i.e. $v_{\text{cond}}(I_i|I_j)$ and $v_{\text{cond}}(I_j|I_i)$.

Given the key-point pairs selected by the correspondence attention module, which forms a undirected graph, we propose a novel discrepancy-based GCN to encode the complex graph structured contextual information into conditional feature vectors.

#### 3.4.1 Discrepancy-based Graph Convolutional Network

In this paper, we follow a most widely used spectral graph convolution [9], which conducts convolution operation in Fourier Domain. Given a feature-map, where each $D$-dimensional feature vector of a pixel is denoted as $x \in R^D$, also known as the node feature in GCN, the graph convolution operation is formulated as follows:

$$g_\theta * x = U g_\theta U^T x$$

where $g_\theta$ is a function of the eigen-value of of the normalized graph Laplacian $L = L - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, $U$ is the matrix of the eigen-vectors of $L$. This equation is computationally expensive due to the need of eigen decomposition of $L$. As a result, a K-order Chebyshev polynomials is used to approximate the convolution operation:

$$g_{\theta'} * x = \sum_{k=0}^{K} \theta_k^* T_k(\tilde{L})x$$

where $\tilde{L} = \frac{2}{\lambda_{\max}} L - I$ with $\lambda_{\max}$ the maximum of the eigen value; $T_kx = 2xT_{k-1}(x) - T_{k-2}(x)$, with $T_0(x) = 1$ and $T_1(x) = x$.

For further simplification, GCN sets $K$ to 1 and $\lambda_{\max}$ is set to a fixed value 2. Thus equation [9] is simplified to:

$$g_{\theta'} * x = \theta_0'x + \theta_1'(L - I)x = \theta_0'x - \theta_1'D^{-\frac{1}{2}} A D^{\frac{1}{2}}x$$

To further decrease the number of learn-able parameter, common GCN sets $\theta = \theta_0' = -\theta_1$, which leads to following expression:

$$g_{\theta} * x = \theta(I + D^{-\frac{1}{2}} A D^{\frac{1}{2}})x$$

As analyzed in [12], by taking a closer look at this equation, we will find that this operation is essentially a weighted average or smooth operation over current node feature itself and its connected neighbours. However, this is not what we want for our graph convolution. Under the setting of the Re-identification, instead of smoothing the value between connected nodes, we require the model to obtain the features difference between the connected key-points. Thus, we propose a novel graph convolution operation, that instead of letting $\theta_0' = -\theta_1$, sets $\theta = \theta_0' = \theta_1'$, which leads to following graph convolution operation:

$$g_{\theta} * x = \theta(I - D^{-\frac{1}{2}} AD^{\frac{1}{2}})x = \theta D^{-\frac{1}{2}} LD^{\frac{1}{2}}x$$
From Eq. 12 we can see that for our new graph convolution, the coefficient of $g_0$ becomes the normalized graph Laplacian matrix, which is equivalent to computing a secondary gradient of the node feature. Hence this convolution is able to obtain the level of feature change between adjacent nodes.

### 3.4.2 Mixed up ID Loss

As shown in Figure 3, given the feature-maps of image pair $(I_i, I_j)$ and the adjacent matrix $A$ generated from the feature-map pair, we obtain the conditional feature-map by applying the graph convolution described in Eq. 12. The outputs are a pair of conditional feature-map with the same size of the input feature-map. Similar to the individual exam stage, the conditional feature-maps are then fed into a feature encoder consisting of a Global Average Pooling layer and a 1 × 1 convolution layer for dimension reduction to obtain the encoded conditional feature vectors.

In the training process, we use both triplet loss and cross entropy loss as the supervised signals. In every training iteration, we sample $P$ identities from the training set and for each identity in the training set, we sample $M$ samples. For triplet loss, we use a common hard triplet loss for person ReID [7].

Training ReID model with merely triplet loss usually causes over-fitting on a very small sub-set of hard samples. Hence an extra cross-entropy loss is required. Since the conditional feature $v_{coup}(I_i|I_j)$ is extracted by GCN based on the information from both of $I_i$ and $I_j$, the identity labels from both image should be used to supervised the feature extraction. Instead of the common cross entropy loss, we propose a mix-up cross entropy loss uses the identity labels of a image pair to train the couple feature vector. Given a mini-batch containing $PM$ images, the mix-up cross-entropy loss is formulated as:

$$L_{mix-up} = \sum_{i=1}^{PM} \sum_{j=1}^{PM} \alpha L_{CE}(y_i, v_{coup}(I_i|I_j))$$

$$+ (1 - \alpha) L_{CE}(y_j, v_{coup}(I_i|I_j))$$

where $L_{CE}$ is the softmax and cross entropy loss shown in Eq.1.

### 3.5. Model Inference

Like most of the ReID method, the model is evaluated as an image retrieval task. Given a query image set containing $N_q$ images and a gallery set containing $N_g$ images, we need to retrieve the images with the same identity of $I_q$. Our method obtains the similarity between two images with both individual features and conditional features.

We first extract the individual features for all images in query and gallery set, with computational complexity of $O(N_q + N_g)$. Then, for each query, we first sort the gallery images based on the similarity of the individual features. After that, the feature-maps of query image and the top-$K$ images in sorted gallery forms $K$ feature-map pairs, which are fed into the key-point alignment stage and conditional feature embedding stage to obtain conditional features, with computational complexity of $O(N_qK)$. Finally, the top-$K$ gallery images are sorted once more by the similarity of coupled features, forming the final ranking result. Compared to the individual feature extraction, much fewer computing operations are needed to obtain conditional feature embedding, so the entire computation cost of Siamese-GCN is very close to normal ReID method.

### 4. Experiments

In this section we propose the performance comparison of Siamese-GCN with the state-of-the-art methods and ablation study of different components in Siamese-GCN.

#### 4.1. Datasets

Our experiments are conducted on three widely used ReID benchmark datasets.

- **Market-1501** [32] dataset contains 32,668 person images of 1,501 identities captured by six cameras. Training set is composed of 12,936 images of 751 identities while testing data is composed of the other images of 750 identities.
- **MSMT-17** [23] dataset contains 124,068 person images of 4,101 identities captured by 15 cameras (12 outdoor, 3 indoor). Training set is composed of 30,248 images of 1,041 identities while testing data is composed of the other images of 3060 identities.
- **DukeMTMC-reID** [15] dataset contains 36,411 person images of 1,404 identities captured by eight cameras. They are randomly divided, with 702 identities as the training set and the remaining 702 identities as the testing set. In the testing set, for each ID in each camera, one image is picked for the query set while the rest remain for the gallery set.

#### 4.2. Implementation Detail

The input images are resized into $384 \times 128$. In training stage, we set batch size to be 16 by sampling 4 identities and 4 images per identity. The ResNet-50 [4] model pretrained on ImageNet is used as the backbone network. Some common data augmentation strategies including horizontal flipping, random cropping, padding and random erasing [33] (with a probability of 0.2) are used. We adopt Gradient Descent optimizer to train our model and set weight decay $5 \times 10^{-4}$. The total number of epoch is 80. The learning rate is initialized to $6.25 \times 10^{-3}$ and is decayed by cosine method until it equals to 0. At the beginning, we warm up the models for 5 epochs and the learning rate grows linearly from 0 to $6.25 \times 10^{-3}$. 

6
4.3. Ablation Study

In this sub-section, we report the evaluation results of the influence of different components and hyper-parameters of our method.

4.3.1 Influence of Model Components

Table 2 shows the influence of the three stages of Siamese-GCN (i.e. individual feature embedding, key-point alignment and conditional feature embedding). Following methods are compared:

- **Individual Feature Embedding.** This is a baseline method using only the individual feature vector extracted by the encoder in the individual exam stage. We observe that this method achieves lowest performance.

- **Key-point Alignment.** The feature-maps extracted by the individual stage are further fed into the Key-point Search stage to get correspondent pixel pairs between two images. Instead of applying GCN, we directly compute the average feature similarity of the cross-image key-point pairs as the overall similarity of the two images. Table 2 shows an extra key-point alignment improves the baseline model by more than 1 percentage point in terms of MAP.

- **Individual-GCN.** All three stages are performed, but all cross image connections in the adjacent matrix are discarded and only intra-frame relations are considered. As show in Table 2, intra-frame relation based GCN gives around 1 percent improvement in terms of MAP, which indicates that extracting features based on local information and correlation inside individual image can boost ReID performance.

- **Siamese-GCN (Normal).** Siamese-GCN considers both intra-frame relation and inter-frame relation in GCN. Here the common GCN that uses a smooth operation on adjacent node is applied and we observe that normal GCN does not achieve performance improvement, which indicates that normal GCN is not suitable for extracting conditional features for ReID.

- **Siamese-GCN (Discrepancy-based).** Our novel discrepancy-based GCN is used and achieves obvious improvement compared to normal GCN. Furthermore, adding inter-frame relation between two images outperforms individual-GCN by a clear margin, showing the effectiveness of extracting conditional features based on local correlation between image pairs.

| Method                                      | mAP   | Rank-1 |
|---------------------------------------------|-------|--------|
| Individual Feature Embedding                | 77.03 | 87.84  |
| + Key-point Alignment and Selection         | 78.29 | 89.68  |
| + Individual-GCN                            | 79.19 | 89.77  |
| + Siamese-GCN (Normal)                      | 78.58 | 89.77  |
| + Siamese-GCN (Discrepancy-based)          | 81.29 | 90.89  |

4.3.2 Influence of Alignment Strategy

In order to prove the advantage of using automatically learned neural network to predict the correspondence between pixel in the feature-map pairs, Table 3 compares the influence of different alignment strategy to our Siamese-GCN method. Following alignment strategy is evaluated:

- **No Alignment:** The baseline method uses only feature vector from individual exam stage where no connection exists between any pixels.

- **Part-based Alignment:** Similar to the part-based model like PCB [20], two pixels at the same location of the images are aligned. As shown in Table 3, part alignment strategy outperforms the baseline method but achieve lower performance than correspondence attention, because it is not robust to scale changes and unable to rule out non-decisive pairs.

- **Fully Connect:** A fully connected adjacent matrix is applied where all pixels are correspondent. In this way, the contextual information for all other pixels are explored when extracting a conditional feature. This method outperforms the baseline but does not achieve the best performance, because too much redundant contextual information is involved in a fully connected graph.

| Method                                      | mAP   | Rank-1 |
|---------------------------------------------|-------|--------|
| Cross Entropy (Baseline)                    | 88.57 | 94.74  |
| Mix-up α = 0.95                            | 89.44 | 95.75  |
| Mix-up α = 0.9                              | 90.30 | 95.96  |
| Mix-up α = 0.8                              | 89.56 | 95.84  |
| Mix-up α = 0.7                              | 88.72 | 95.43  |
| Mix-up α = 0.6                              | 85.67 | 94.83  |

| Method                                      | mAP   | Rank-1 |
|---------------------------------------------|-------|--------|
| No Alignment (Baseline)                     | 77.03 | 87.84  |
| Part Alignment                              | 78.39 | 89.45  |
| Fully Connect                               | 80.19 | 89.90  |
| Top-K Similarity                            | 79.30 | 89.72  |
| Correspondence Attention                     | 81.29 | 90.89  |

Table 2. Performance (%) comparisons of three stages (individual exam, key-point search and joint exam) on DukeMTMC.

Table 3. Performance (%) comparisons of different alignment strategy for graph generation in Siamese-GCN on DukeMTMC.

Table 4. Performance (%) comparisons of different alignment strategy for graph generation in Siamese-GCN on Market1501.
Table 5. Performance (%) comparisons to the state-of-the-art results on Market-1501, DukeMTMC-reID and MSMT-17. Our proposed Siamese-GCN outperforms the state-of-the-art methods.

| Category | Method             | Market-1501 | DukeMTMC-reID | MSMT-17   |
|----------|--------------------|-------------|---------------|-----------|
|          |                    | mAP | Rank-1 | mAP | Rank-1 | mAP | Rank-1 |
| Part-based | PCB [20]           | 77.4 | 92.3 | 66.1 | 81.7 | - | - |
|          | MGN [21]           | 86.9 | 95.7 | 78.4 | 88.7 | - | - |
|          | Pyramid [30]      | 88.2 | 95.7 | 79.0 | 89.0 | - | - |
|          | OSNet [34]         | 84.9 | 94.8 | 73.5 | 88.6 | 52.9 | 78.7 |
|          | SPReID [8]         | 83.36 | 93.68 | 73.34 | 85.95 | - | - |
|          | DSA-reID [27]       | 87.6 | 95.7 | 74.3 | 86.2 | - | - |
| Alignment | PCB+RPP [20]       | 81.6 | 93.8 | 69.2 | 83.3 | - | - |
|          | AlignedReID [20]   | 79.3 | 91.8 | - | - | - | - |
|          | FD-GAN [11]        | 77.7 | 90.5 | 64.5 | 80.0 | - | - |
|          | VPM [19]           | 80.8 | 93.0 | - | - | - | - |
| Attention | HA-CNN [13]        | 75.7 | 91.2 | 63.8 | 80.5 | - | - |
|          | ABD-Net [2]        | 88.28 | 95.60 | 78.59 | 89.00 | 60.80 | 82.30 |
| Joint Learning | DCCs [24]       | 71.1 | 88.4 | 59.2 | 80.3 | - | - |
|          | SMI [25]           | 65.25 | 86.15 | - | - | - | - |
| Graph-based | Group-shuffling [16]   | 82.5 | 92.7 | 66.4 | 80.7 | - | - |
|          | SGGNN [17]         | 82.8 | 92.3 | 68.2 | 81.1 | - | - |
| This work | Siamese-GCN         | 90.30 | 95.96 | 81.29 | 90.89 | 62.00 | 83.54 |

Figure 4. The cross-image alignment result obtained by similarity-based alignment and correspondence attention. The red lines denote a key-point correspondence between two images.

- **Similarity-based Alignment**: Similar to AlignedReID [26], this method selects the most similar pixel as each pixel’s correspondent neighbour. It does not perform as well as our method because as discussed in section 1, the predefined similarity-based alignment strategy is not flexible enough for various types of image pairs.

- **Correspondence Attention Module**: With a correspondence attention module, Siamese-GCN is able to build a more flexible correspondence compared to pre-defined rules and achieves the best performance.

In Figure 4 we compare the visualization results of similarity-based alignment and correspondence attention, where key-point pairs with highest scores are visualized. We observe that similarity-based alignment only aligns key-point pairs with high visual similarity, causing mismatching images with similar local parts (e.g., the lower body in Figure 4 a,b and upper-body in Figure 4 c). On the other hand, our correspondence attention disregards the visual similarity and is able to focus on correct decisive key-points to reject image pairs (e.g. the shoulder area in Figure 4).

4.3.3 Influence of Mix-up ID Loss

Table 4 shows the influence of our customized Mix-up Loss for conditional feature embedding. As shown in Table 4 with the right hyper-parameter \( \alpha \), our proposed Mix-up Loss significantly outperforms common cross entropy loss. We also observe the influence of the hyper-parameter \( \alpha \) to the model performance, where Siamese-GCN achieves the best performance when \( \alpha \) is set to 0.9, which shows that conditional feature contains only small amount of information from contextual image compared to the target image.
4.4. Comparison with the State-of-the-Art

We evaluate our proposed Siamese-GCN with the state-of-the-art ReID models. These methods include: (1) the part-based models such as MGN, Pyramid; (2) the alignment-based methods like PCB, AlignReID; (3) the human semantic parsing-based methods like SPReID, DSA-reID; (4) the attention-based methods like HA-CNN, ABD-Net, Robust ReID; (5) Joint learning methods including DCCs and SMI that learns conditional features with RNN; (6) ReID methods that utilizes graph structure such as Group-shuffling Random Walk and SGGNN. Table 4.3.1 shows the performance comparison of Siamese-GCN with State-of-the-Art methods. As shown in Table 4.3.1, thanks to our novel ReID framework that integrates key-point alignment and conditional feature embedding, Siamese-GCN outperforms the state-of-the-art methods on Market1501, DukeMTMC and MSMT-17.

5. Conclusions

In this paper we propose a novel Person ReID framework that integrates both key-points alignment and conditional feature embedding. Our proposed Siamese-GCN is able to automatically select decisive key-point pairs with correspondence attention module, and extract conditional feature embedding from the key-point pairs with a novel discrepancy-based GCN. The experiments show the effectiveness of our model.

References

[1] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:1312.6203, 2013.

[2] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. arXiv preprint arXiv:1908.01114, 2019.

[3] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, et al. Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In Advances in neural information processing systems, pages 1222–1233, 2018.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[5] Lingxiao He, Jian Liang, Haiqing Li, and Zhenan Sun. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7073–7082, 2018.

[6] Lingxiao He, Zhenan Sun, Yuhao Zhu, and Yunbo Wang. Recognizing partial biometric patterns. arXiv preprint arXiv:1810.07399, 2018.

[7] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737, 2017.

[8] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gkmen, Mustafa E. Kamasak, and Mubarak Shah. Human semantic parsing for person re-identification. In CVPR, June 2018.

[9] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907, 2016.

[10] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 384–393, 2017.

[11] Jianing Li, Shiliang Zhang, Qi Tian, Meng Wang, and Wen Gao. Pose-guided representation learning for person re-identification. IEEE transactions on pattern analysis and machine intelligence, 2019.

[12] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[13] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In CVPR, June 2018.

[14] Jiaxu Miao, Yu Wu, Ping Liu, Yuhang Ding, and Yi Yang. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 542–551, 2019.

[15] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In European Conference on Computer Vision, pages 17–35. Springer, 2016.

[16] Yantao Shen, Hongsheng Li, Tong Xiao, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Deep group-shuffling random walk for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2265–2274, 2018.

[17] Yantao Shen, Hongsheng Li, Shuai Yi, Dapeng Chen, and Xiaogang Wang. Person re-identification with
deep similarity-guided graph neural network. In Proceedings of the European conference on computer vision (ECCV), pages 486–504, 2018.

[18] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. Dual attention matching network for context-aware feature sequence based person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5363–5372, 2018.

[19] Yifan Sun, Qin Xu, Yali Li, Chi Zhang, Yikang Li, Shengjin Wang, and Jian Sun. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 393–402, 2019.

[20] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling. In ECCV, September 2018.

[21] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi Zhou. Learning discriminative features with multiple granularities for person re-identification. In ACM MM, pages 274–282. ACM, 2018.

[22] Kan Wang, Changxing Ding, Stephen J Maybank, and Dacheng Tao. Cdpm: convolutional deformable part models for semantically aligned person re-identification. IEEE Transactions on Image Processing, 2019.

[23] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 79–88, 2018.

[24] Lin Wu, Yang Wang, Junbin Gao, and Dacheng Tao. Deep co-attention based comparators for relative representation learning in person re-identification. arXiv preprint arXiv:1804.11027, 2018.

[25] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. Pattern Recognition, 76:727–738, 2018.

[26] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184, 2017.

[27] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In CVPR, June 2019.

[28] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1077–1085, 2017.

[29] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In The IEEE International Conference on Computer Vision, Oct 2017.

[30] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In CVPR, June 2019.

[31] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose invariant embedding for deep person re-identification. IEEE Transactions on Image Processing, 2019.

[32] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In The IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[33] Chun Zhong, Liang Zheng, Guoliang Kang, Shaoz Li, and Yi Yang. Random erasing data augmentation. arXiv preprint arXiv:1708.04896, 2017.

[34] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, pages 3702–3712, 2019.

[35] Zhihui Zhu, Xinyang Jiang, Feng Zheng, Xiaowei Guo, Feiyue Huang, Weishi Zheng, and Xing Sun. Viewpoint-aware loss with angular regularization for person re-identification. arXiv preprint arXiv:1912.01300, 2019.