Group Consistent Similarity Learning via Deep CRF for Person Re-Identification

Dapeng Chen[1,3] Dan Xu[2] Hongsheng Li[1†] Nicu Sebe[2] Xiaogang Wang[1†]
[1]CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
[2]Department of Information Engineering and Computer Science, University of Trento
[3]School of Software Engineering, Xi'an Jiaotong University
{dpchen, hsli, xgwang}@ee.cuhk.edu.hk {dan.xu, niculae.sebe}@unitn.it

Abstract

Person re-identification benefits greatly from deep neural networks (DNN) to learn accurate similarity metrics and robust feature embeddings. However, most of the current methods impose only local constraints for similarity learning. In this paper, we incorporate constraints on large image groups by combining the CRF with deep neural networks. The proposed method aims to learn the "local similarity" metrics for image pairs while taking into account the dependencies from all the images in a group, forming "group similarities". Our method involves multiple images to model the relationships among the local and global similarities in a unified CRF during training, while combines multi-scale local similarities as the predicted similarity in testing. We adopt an approximate inference scheme for estimating the group similarity, enabling end-to-end training. Extensive experiments demonstrate the effectiveness of our model that combines DNN and CRF for learning robust multi-scale local similarities. The overall results outperform those by state-of-the-arts with considerable margins on three widely-used benchmarks.

1. Introduction

Person re-identification (Re-ID) is a critical task in intelligent video surveillance, aiming to associate the same people across different cameras. It is generally formulated as a ranking problem: given a probe image of a person, the algorithm needs to rank all gallery images based on their similarities w.r.t. the probe image. The ranking performance heavily relies on the quality of similarity metric, which is usually learned from the data.

Encouraged by the remarkable success of deep neural networks (DNN), the Re-ID community also employs DNNs for end-to-end similarity learning. A common practice is to employ local constraints. For instance, most methods straightforwardly make use of the pairwise constraints between image samples [1, 38, 40] (Fig. 1a), trying to minimize the distances between positive pairs while maximizing the distances between negative pairs. Beyond the pairwise constraints, several methods adopt the triplet loss [9, 48, 14] to enforce a correct ranking order (Fig. 1b). Recently, a quadruplet loss [7] is proposed to further improve the triplet loss by reducing the intra-class variations and enlarging the inter-class variations (Fig 1c). To make use of these local constraints with DNNs, existing approaches have to sample small cliques such as pairs, triplets or quadruplets, which are further used to organize training batches and construct the optimization losses, making the learning of similarity metric largely dependent on the sampling strategies. As most of the local constraints can be easily satisfied by the learned similarity metrics during training, local constraints are less efficient to contribute useful learning signals. Furthermore, with stochastic gradient descent method, the con-

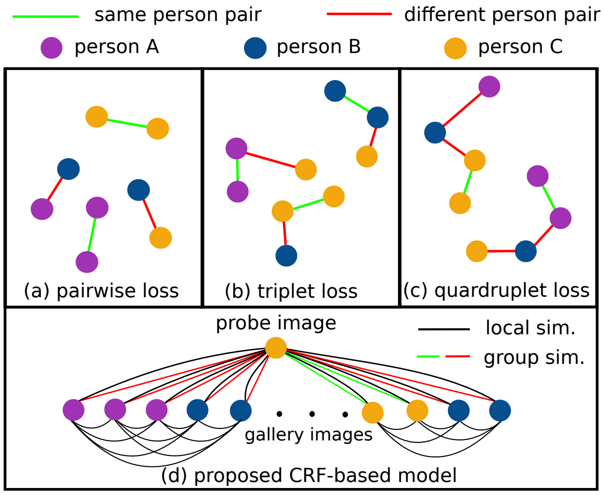


Figure 1: Illustration of different constraints for similarity learning. (a) pairwise loss, (b) triplet loss, (c) quadruplet loss, and (d) proposed CRF based model. The green lines connect positive pairs whose in-between distances need to be minimized, while red lines indicate the negative pairs whose in-between distances need to be maximized.

[†] H. Li and X. Wang are the co-corresponding authors.

straints optimized by one update can probably become invalid by another update, leading to suboptimal solutions.

Instead of imposing local constraints over small clips, we propose to leverage supervision with image groups and model more complex image-to-image relations. Each group consists of a probe image and a set of gallery images. We define "local similarity" and "global similarity" to describe the inter-image relationships, which are based on the related two images and the whole image group, respectively. The two kinds of similarities are associated in a unified graphical model by a Conditional Random Field (CRF), where the local similarities are input variables that have been observed while the group similarities are output variables to be predicted. As diverse dependencies are modeled in the CRF, optimizing the group similarities can in turn learn more consistent local similarity metrics as well as feature embeddings. Besides, benefited from the flexibility and representative power of the graphical model, we can effectively fuse different types of local similarities of multi-scale feature embeddings for more accurate similarity estimation. To implement our model with DNN, we derive approximate inference to estimate the group similarity, yielding mean-field updating procedure. Three network modules are designed for multi-scale feature embeddings (MFE), local similarity computation (LS) and group similarity estimation (GS), respectively. It is noteworthy that we only perform group similarity estimation in the training stage. The similarity to be predicted in testing is the linear combination of multi-scale local similarities, where the combination parameters are adaptively learned from the CRF.

In summary, our main contributions are as follows. (1) We combine the CRF model with DNN to learn more consistent multi-scale similarity metrics. Various inter-image dependencies within an image group are modeled by a unified graphic model. (2) We adopt approximate inference scheme for our model and implement the inference procedure via neural network modules, allowing end-to-end training. (3) Extensive ablation studies validate the effectiveness of employing group similarities within the CRF for training. It benefits the feature embeddings, local similarities and multi-scale similarity combination. We evaluate our approach on three large-scale Re-ID datasets and the results outperform those by state-of-the-art methods.

2. Related Work

Early works on person Re-ID concentrated on either feature extraction [42, 27, 10, 12] or metric learning [16, 22, 3, 29]. Recent methods mainly benefit from the advances of CNN architectures, which learn the two aspects in an end-to-end fashion [20, 1, 4, 38, 40, 5, 19, 7]. Our work can be uniquely positioned as deep similarity learning with CRF.

A typical category of deep similarity learning for person Re-ID is to train a siamese network with contrastive loss [38, 39, 40, 1], where the task is to reduce the distances between images of the same person and to enlarge the distances between the images of different persons. One downside of this approach is that it focuses on absolute distances, whereas relative distances are more important for a ranking problem like person Re-ID. Several methods [11, 9, 48] employed triplet loss to enforce the correct order of relative distances among image triplets, *i.e.*, the positive image pair is more similar than the negative image pair w.r.t. a same anchor image. Chen *et al.* [7] proposed quadruplet loss which combined the the advantages of contrastive loss and triplet loss, complementing the triplet loss by minimizing the intra-class variations and maximizing the inter-class variations. However, all these constraints are based on small clips which do not take the global structure of the embedding space into consideration, usually leading inefficient training and sub-optimal solutions. In fact, recent works began to develop effective sampling strategies [33, 14, 43, 1]. They evidently improve the local constraints often relies on expensive computational requirements and may be sensitive to data distribution.

To overcome the limitation of local constraints, we adopt the CRF model [18] to connect various dependencies within a large image group, and combine it with DNN. Our approach is motivated by the advances in semantic image segmentation [50, 6, 24] and depth/surface normals estimation [45, 41], where they implement the mean-field inference for CRF [17, 32] in an end-to-end learnable neural networks. However, different from these methods that build the inter-pixel dependencies in a single image, our approach models the inter-image dependencies in a training batch. Deep metric learning methods [30, 35] also stressed incorporating more images in the training constraints. They either mine the hard negative samples or enforce clustering for the images with the same label, while our model associates all the group images in a unified graphical model, aiming to learn more consistent similarity metrics within the group. During training, we apply the verification loss on the group similarities and employ the identification loss on the feature embeddings to supervise the similarity learning. The effectiveness of such joint identification-verification losses have been validated by [36, 40, 51, 21], which generally adopted cross-entropy loss for the identification of the feature embeddings. Slightly different from their method, we choose the OIM loss proposed in [44] for the identification, which is scalable to large dataset where each person can have a variable number of person images.

3. Our Approach

Our method aims to learn more robust similarity metric for the Re-ID task by taking into account the inter-image relations within image groups. We define "local similarities" and "global similarities" in the image group (Section 3.1),

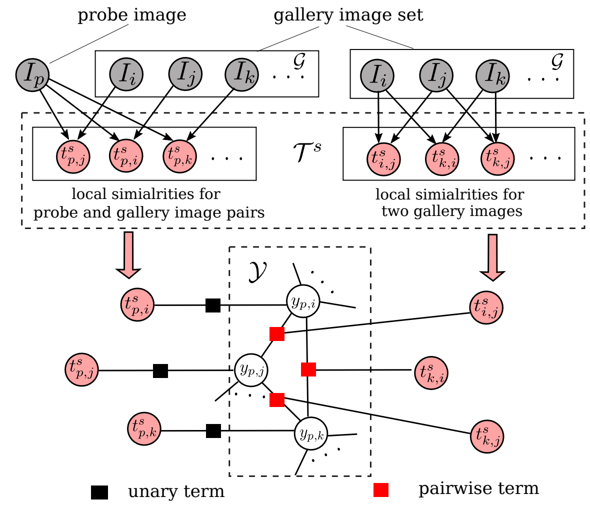


Figure 2: Graphical model for the local similarities and group similarities in an image group. The local similarities $\mathcal{T}$ between image pairs are estimated by a deep neural network. The group similarities $\mathcal{Y}$ are random variables conditioned on $\mathcal{T}$, and the distribution $P(\mathcal{Y}|\mathcal{T})$ is modeled as the CRF. We only illustrate the local similarities with a single scale s.

then jointly model them in the CRF (Section 3.2). For network implementation, we adopt an approximate inference scheme for group similarity estimation (Section 3.3), and design three network modules (Section 3.4). The overall network and training details are introduced in Section 3.5 and Section 3.6.

3.1. Local Similarity and Group Similarity

The training data for DNNs are usually organized in batches, where the images not only participate in a same forward and backward pass for optimization, but also contain abundant inter-images relations that allow us to exploit. We construct image groups in a batch, and each of them can have a flexible number of images.

Let $\mathcal{O}$ denote all the images in a group. Among them, there is a probe image I_p, and the remaining images are gallery images forming the set $\mathcal{G} = \{I_1, I_2, ..., I_G\}$. We define $t_{m,n}$ to represent the "local similarity" for two arbitrary images I_m and I_n in the group, and $t_{m,n}$ is only related to the appearance of the two images. On the other hand, the similarity between two images can also be inferred by their relations to other images. We further define "group similarity" $y_{p,i}$ between the probe image I_p and an arbitrary gallery image I_i in $\mathcal{G}$. $y_{p,i}$ makes use of the whole image group for similarity estimation.

In this work, both local similarity $t_{m,n}$ and group similarity $y_{p,i}$ are assumed to be within the range $(0, 1)$. The higher values the similarities are, the more likely the two images belong to a same person.

3.2. Group Consistency Modeling via CRF

Given a group of images $\mathcal{O}$, we first estimate the local similarities $\mathcal{T}$ for the image group. In particular, we consider multi-scale local similarities and each local similarity are about two arbitrary images in the group. Therefore, we have $\mathcal{T} = \{\mathcal{T}^s\}_{s=1}^S$ and $\mathcal{T}^s = \{t^s_{m,n}|I_m, I_n \in \mathcal{O}\}$, where $\mathcal{T}^s$ contains the local similarities of scale s. In this work, the local similarity $t^s_{m,n}$ is computed via a deep neural network, denoted by a function:

$$t^s_{m,n} = \xi^s(I_m, I_n), \tag{1}$$

where $\xi^s(I_m, I_n)$ computes the similarity based on the scale s feature embeddings $\phi^s(I_m)$ and $\phi^s(I_n)$.

The group similarities are modeled as random variables that describe the similarities between the probe image and gallery images, forming the set $\mathcal{Y} = \{y_{p,i}|I_i \in \mathcal{G}\}$. They are conditioned on the local similarities $\mathcal{T}$, and the pair $(\mathcal{Y}, \mathcal{T})$ can be modeled as the continuous CRF, characterized by a Gibbs distribution:

$$P(\mathcal{Y}|\mathcal{T}) = \frac{1}{Z(\mathcal{T})} \exp(-E(\mathcal{Y}|\mathcal{T})), \tag{2}$$

where $Z(\mathcal{T})$ is the partition function and $E(\mathcal{Y}|\mathcal{T})$ is the energy function. For the fully connected pairwise CRF model, $E(\mathcal{Y}|\mathcal{T})$ can be represented as:

$$\sum_{s=1}^{S} \Big(\alpha^s \sum_i \Psi_u(y_{p,i}, t^s_{p,i}) + \beta^s \sum_{i<j} \Psi_p(y_{p,i}, y_{p,j}, t^s_{i,j})\Big), \tag{3}$$

where α^s and β^s are positive parameters associated with the unary terms and pairwise terms of scale s. With these terms, the energy function models the relations between the multi-scale local similarities and group similarities. More specifically, the unary term is given by:

$$\Psi_u(y_{p,i}, t^s_{p,i}) = (y_{p,i} - t^s_{p,i})^2. \tag{4}$$

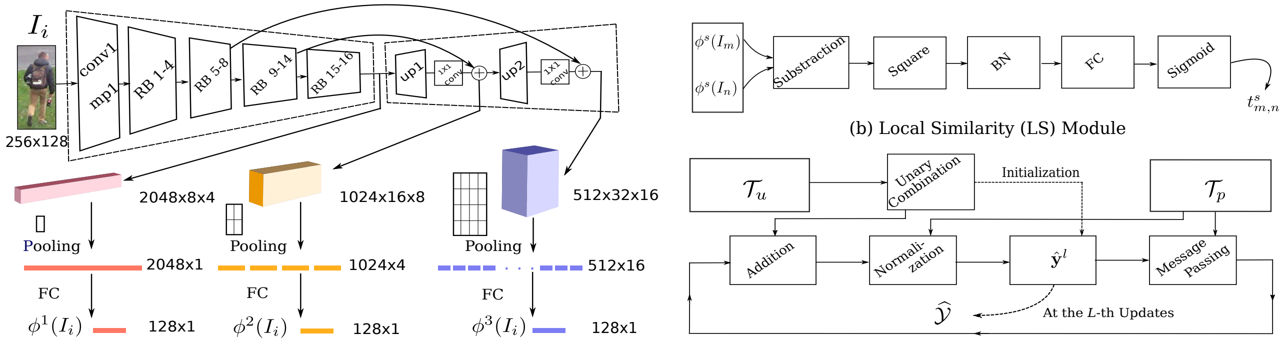
It enforces the group similarity $y_{p,i}$ to be close to the local similarity $t^s_{p,i}$, which predicts the group similarity without considering the consistency of other images in the group. The pairwise term is :

$$\Psi_p(y_{p,i}, y_{p,j}, t^s_{i,j}) = t^s_{i,j}(y_{p,i} - y_{p,j})^2. \tag{5}$$

If the local similarity $t^s_{i,j}$ of I_i and I_j is high, the two images are encouraged to be commonly similar or dissimilar to the probe image I_p. Such assumption enhances the consistency among the group similarities between gallery images. The graphical model for the proposed CRF is depicted in Fig. 2.

3.3. Approximate Inference

After obtaining local similarities $\mathcal{T}$, we exploit the mean-field approximation to derive a tractable inference procedure. It approximates $P(\mathcal{Y}|\mathcal{T})$ by a simpler distribution



(a) Multi-scale Feature Embedding (MFE) Module

(b) Local Similarity (LS) Module

(c) Group Similarity (GS) Module

Figure 3: Network Structures for different modules. (a), learning multi-scale feature embeddings with ResNet-50 as the backbone network, where RB represents the residual block. (b)-(c) are the network modules for estimating local similarities and group similarities, where $\mathcal{T}_u$ contains the local similarities for unary term, $\mathcal{T}_p$ contains the local similarities for pairwise term and $\hat{\mathbf{y}}^l$ is the vector containing the group similarities at the lth iteration.

$Q(\mathcal{Y})$, which can be written as the product of a set of independent marginal distributions, *i.e.*, $Q(\mathcal{Y}) = \prod_i Q_i(y_{p,i})$. By minimizing the KL divergence between $P(\mathcal{Y}|\mathcal{T})$ and $Q(\mathcal{Y})$[2], the optimal distribution $\widehat{Q}_i(y_{p,i})$ is estimated by:

$$\ln \widehat{Q}_i(y_{p,i}) = \mathbb{E}_{j\neq i}[\ln P(\mathcal{Y}|\mathcal{T})] + \text{const}, \tag{6}$$

where $\mathbb{E}_{j\neq i}[\cdot]$ denotes an expectation under $Q(\mathcal{Y})$ over all group similarities except $y_{p,i}$. By expanding $P(\mathcal{Y}|\mathcal{T})$ with Eq. (3), $\widehat{Q}_i(y_{p,i})$ can be written as:

$$\begin{aligned}\widehat{Q}_i(y_{p,i}) \propto \exp\Big(&\sum_{s=1}^{S}(\alpha^s \Psi_u(y_{p,i}, t^s_{p,i}) \\ &+ \beta^s \sum_{j\neq i} \mathbb{E}[\Psi_p(y_{p,i}, y_{p,j}, t^s_{i,j})])\Big).\end{aligned} \tag{7}$$

The definitions of $\Psi_u(y_{p,i}, t^s_{p,i})$ and $\Psi_p(y_{p,i}, y_{p,j}, t^s_{i,j})$ imply that $\widehat{Q}_i(y_{p,i})$ is a Gaussian function, whose expectation also yields the maximum probability, denoted by $\hat{y}_{p,i}$. By taking Eq. (4) and Eq. (5) into $\widehat{Q}_i(y_{p,i})$, we can have the followings updates for $\hat{y}_{p,i}$:

$$\hat{y}^{l+1}_{p,i} = \frac{\sum_{s=1}^{S} \alpha^s t^s_{p,i} + \sum_{s=1}^{S} \beta^s \sum_{j\neq i} t^s_{i,j} \hat{y}^l_{p,j}}{\sum_{s=1}^{S} \alpha^s + \sum_{s=1}^{S} \beta^s \sum_{j\neq i} t^s_{i,j}}, \tag{8}$$

where one group similarity is influenced by both the local similarities and group similarities. As the mean-field algorithm will generally achieve convergence after L iterations, the final estimated group similarity $\hat{y}_{p,i} = \hat{y}^L_{p,i}$. We collect the estimated group similarities in set: $\widehat{\mathcal{Y}} = \{\hat{y}_{p,i}|I_i \in \mathcal{G}\}$.

3.4. CRF modeling with Deep Neural Network

The implementation of our deep CRF model consists of the modules for multi-scale feature embedding, local similarity estimation and the group similarity estimation.

Feature embedding module (Fig. 3a). The local and group similarities are calculated based on multi-scale feature maps generated by a DNN. Inspired by Feature Pyramid Network (FPN) [25], the feature embedding module takes ResNet-50 [13] as backbone, and generates the multi-scale high-level semantic feature maps by combining the top-down pathway and lateral connections. In particular, the top-down pathway employs upsampling and 1×1 convolutions to match the lateral input in both spatial dimension and feature dimension. To obtain the feature maps for the whole image, we apply non-overlapped 8×4 spatial pooling to feature maps of all scales. Such pooling strategy can partially preserve the spatial structure of the feature maps at larger scales, and balance the semantic and spatial information.

Local similarity module (Fig. 3b). The local similarity $t^s_{m,n}$ is estimated based on $\phi^s(I_m)$ and $\phi^s(I_n)$, the s scale feature embeddings of images I_m and I_m. More specifically, we compute the difference vector of the two feature embeddings, perform an element-wise square operation over the vector, and normalize the vector by a BN layer [15]. The resulting vector is mapped to a scalar via a fully-connected layer, which is further normalized to $(0, 1)$ via a sigmoid function, indicating the probability of I_m and I_n belonging to the same person.

Group similarity module (Fig. 3c). The group similarities $\mathcal{Y}$ are conditioned on the local similarities $\mathcal{T}$, which can be further divided into $\mathcal{T}_u$ and $\mathcal{T}_p$ to be used in unary terms and pairwise terms, respectively. Among them, $\mathcal{T}_u$ contains the local similarities between the probe and gallery images, *i.e.*, $\mathcal{T}_u = \{t^s_{p,k}|I_k \in \mathcal{G}, s = 1, 2, ..., S\}$, while $\mathcal{T}_p$ contains the local similarities between all pairs of gallery images, *i.e.*, $\mathcal{T}_p = \{t^s_{i,j}|I_i, I_j \in \mathcal{G}, s = 1, 2, ..., S\}$. The parameters $\{\alpha^s\}^S_{s=1}$ and $\{\beta^s\}^S_{s=1}$ are required to be positive, we generate them by exponential mappings of trainable parameters, *i.e.*, $\alpha^s = \exp(w^s)$ and $\beta^s = \exp(v^s)$, and initialize $\{w^s\}^S_{s=1}$ and $\{v^s\}^S_{s=1}$ to be zeros. With $\{\alpha^s\}^S_{s=1}$, we further initialize $\hat{y}^0_{p,i}$ by $\sum^S_{s=1} \alpha^s t^s_{p,i} / \sum^S_{s=1} \alpha^s$.

According to Eq. (8), the updating of group similarity consists of several steps. (i) Unary combination, which computes the information from unary terms, *e.g.*,

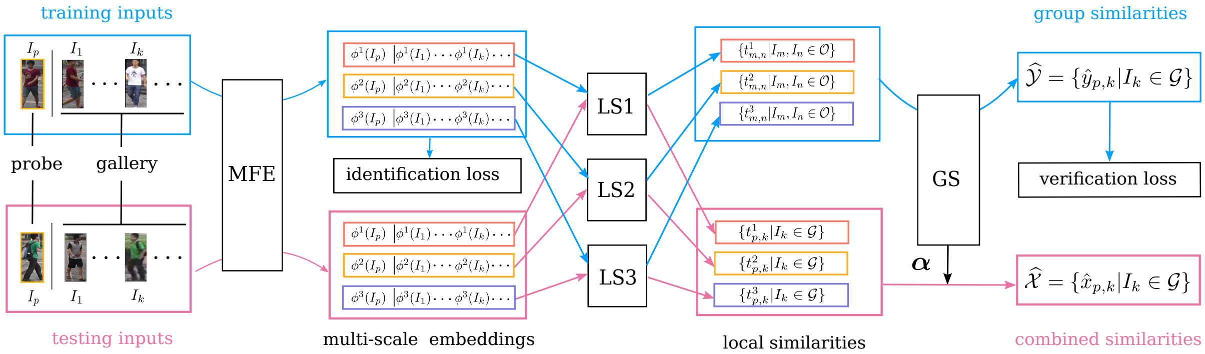

training inputs
I_p
I_1
I_k
probe
gallery
I_p
I_1
I_k
testing inputs
MFE
$\phi^1(I_p)$ $\phi^1(I_1)\cdots\phi^1(I_k)\cdots$
$\phi^2(I_p)$ $\phi^2(I_1)\cdots\phi^2(I_k)\cdots$
$\phi^3(I_p)$ $\phi^3(I_1)\cdots\phi^3(I_k)\cdots$
identification loss
$\phi^1(I_p)$ $\phi^1(I_1)\cdots\phi^1(I_k)\cdots$
$\phi^2(I_p)$ $\phi^2(I_1)\cdots\phi^2(I_k)\cdots$
$\phi^3(I_p)$ $\phi^3(I_1)\cdots\phi^3(I_k)\cdots$
multi-scale embeddings
LS1
LS2
LS3
$\{t^1_{m,n}|I_m, I_n \in \mathcal{O}\}$
$\{t^2_{m,n}|I_m, I_n \in \mathcal{O}\}$
$\{t^3_{m,n}|I_m, I_n \in \mathcal{O}\}$
$\{t^1_{p,k}|I_k \in \mathcal{G}\}$
$\{t^2_{p,k}|I_k \in \mathcal{G}\}$
$\{t^3_{p,k}|I_k \in \mathcal{G}\}$
local similarities
GS
$\boldsymbol{\alpha}$
group similarities
$\widehat{\mathcal{Y}} = \{\hat{y}_{p,k}|I_k \in \mathcal{G}\}$
verification loss
$\widehat{\mathcal{X}} = \{\hat{x}_{p,k}|I_k \in \mathcal{G}\}$
combined similarities

Figure 4: Network architectures for training and testing. The blue flowchart indicates the training network, and red flowchart indicates the testing network.

$\sum_{s=1}^{S} \alpha^s t^s_{p,i}$ for $\hat{y}^{l+1}_{p,i}$, by using the local similarities in $\mathcal{T}_u$. (ii) Message passing, which computes the passed messages, *e.g.*, $\sum_{s=1}^{S} \beta^s \sum_{j\neq i} t^s_{i,j}\hat{y}^l_{p,j}$ for $\hat{y}^{l+1}_{p,i}$, by using the local similarities in $\mathcal{T}_p$. We can compute the messages for all the group similarities by matrix multiplication $\mathbf{T}\hat{\mathbf{y}}^l$. Taking i, j as the ordered image indexes in $\mathcal{G}$ that consists of G images, $\mathbf{T} \in \mathbb{R}^{G\times G}$ is a fixed symmetric matrix:

$$\mathbf{T}_{ij} = \begin{cases} \sum_{s=1}^{S} \beta^s t^s_{i,j}, & \text{if} \quad i \neq j \\ 0, & \text{if} \quad i = j \end{cases}, \quad I_i, I_j \in \mathcal{G}, \tag{9}$$

and $\hat{\mathbf{y}}^l$ is a vector composed by all the group similarities estimated from the last update, *i.e.*, $\hat{\mathbf{y}}^l = [\hat{y}^l_{p,1}, \hat{y}^l_{p,2}, ..., \hat{y}^l_{p,G}]^\top$. (iii) Normalization, which first calculates the normalization factor and then performs the normalization with element-wise division. After L iterations, we put the elements in vector $\hat{\mathbf{y}}^L$ into $\widehat{\mathcal{Y}}$ as the estimation for group similarities. According to Eq. (8) and $t^s_{p,i} \in (0, 1)$, it is easy to prove that $\hat{y}_{p,i} \in (0, 1)$, which still can be used to represent the probability of being a same person.

3.5. Overall Network Architecture

The network architectures for training and testing are demonstrated in Fig. 4. In training, the inputs of training network are in the form of image groups, each of which consists of a probe image and multiple gallery images. The group similarities play two roles: (1) they guide the learning of local similarity metrics considering the diverse dependencies in the group; (2) they learn the linear weights to combine the multi-scale local similarities for more accurate estimation. In testing, the network inputs can be an arbitrary number of probe image and gallery images, and the final similarity is the linear combination of multi-scale local similarities with $\{\alpha^s\}_{s=1}^S$ learned from the GS module:

$$\hat{x}_{p,k} = \sum_{s=1}^{S} \alpha^s t^s_{p,k} / \sum_{s=1}^{S} \alpha^s. \tag{10}$$

Discussion: One important reason that prevents us from adopting group similarities for prediction is the inconsistency between the training and testing configurations. In training, we build a fully connected graph between group similarities (see Fig. 2). In testing, the corresponding graph structure are much larger as there are more gallery images. The differences between the graphs make the learned message passing parameters $\{\beta^s\}_{s=1}^S$ cannot be directly applicable for the testing data. As local similarity metrics have been benefited from learning with group similarity, the predicted similarity (Eq. (10)) can further combine multi-scale local similarities and is flexible to be applied to different testing configurations. To some extent, the proposed CRF model can also be regarded a special loss function with trainable parameters, which mediates among more abundant inter-image constraints.

3.6. End-to-end Optimization

Batch organization. In our implementation, an image batch $\mathcal{B}$ contains images of $N_\mathcal{B}$ person identities, and each person identity has $K_\mathcal{B}$ images. With the image batch, we can form $N_\mathcal{B}$ groups with each group having a probe image from a different identity, the remaining images are gallery images shared by all the groups. In this way, local similarities between two gallery images can be reused by different groups, largely reducing the computational cost.

Loss functions. Since the group similarity represents the same-person probability of an image pair, we can apply the binary cross-entropy loss to each image pair, treating the similarity learning as a verification problem:

$$L^B_{veri}(I_p, I_k) = \begin{cases} -\lambda \log(\hat{y}_{p,k}) & \text{if} \quad l_{p,k} = 1 \\ -(1-\lambda)\log(1-\hat{y}_{p,k}) & \text{if} \quad l_{p,k} = 0 \end{cases} \tag{11}$$

where the label $l_{p,k} = 1$ if the probe image I_p and the gallery image I_k belong to the same person, otherwise $l_{p,k} = 0$, λ is a hyper-parameter to adjust the importance of positive and negative image pairs. The MFE also predicts the person identities during the training. We employ the OIM loss [44] to supervise the per-image multi-scale feature embeddings:

$$L^B_{id}(I_k) = -\sum_{i=1}^{N_{tr}} \sum_{s=1}^{S} l'_{k,i} \log \left(\frac{\exp(\mathbf{w}^s_i \cdot \phi^s(I_k))}{\sum_{j=1}^{N_{tr}} \exp(\mathbf{w}^s_j \cdot \phi^s(I_k))} \right). \tag{12}$$

There are totally N_{tr} identities in the training set, if the image I_k belongs to the ith identity, $l'_{k,i} = 1$, otherwise

$l'_{k,i} = 0$. $\mathbf{w}_i^s$ are the coefficients associated with the s scale feature embedding of the ith identity. They are obtained by using an online updated buffer and measuring similarities between the current person and all other persons in the feature buffer with inner product. The final loss function for each batch is a linear combination of the verification loss averaged over all the group similarities and the identification loss averaged over all the images.

4. Experiments

We evaluate the proposed approach on three datasets. Ablation studies are mainly conducted on Market-1501 [49] and DukeMTMC-reID [52], which have fixed training / testing splits and thus are convenient for extensive evaluation. We also report the final results on CUHK03 [20] to compare with other methods in addition to the above two datasets.

4.1. Experimental Setup

Datasets. All the employed datasets contain multiple images for each person identity. Among them, Market-1501 consists of 32,668 image, including 12,936 training images from 751 identities and 19,732 testing images from 750 identities. DukeMTMC-reID is a subset of the multi-target, multi-camera pedestrian tracking dataset [31]. It contains 1,812 identities captured by 8 cameras. There are 36,411 images in total, where 16,522 images of 702 identities are used for training, 2,228 images of another 702 identities are used as query images, and the remaining 17,661 images are gallery images. In our experiments, we follow the standard single-query protocol [49] for both Market-1501 and DukeMTMC-reID. CUHK03 contains 13,164 images of 1,467 identities. We follow the standard single-shot protocol for the labeled images and detected images separately, which needs to repeat 20 times of random 1,367/100 training/testing identity splitting and report the averaged results.

Implementation details. For our implementation, the input images are resized to 256 × 128 after random cropping and flipping, and REDA [54] is used for data augmentation. Stochastic gradient descent is applied with a momentum of 0.9. The initial learning rate is 0.01, which is further decayed to 0.001 after the 50th epochs. The iteration number L is set to 6. The weighting factor λ in Eq. (11) is set to be 0.7. Each batch contains $N_{\mathcal{B}} = 15$ persons and each persons has $K_{\mathcal{B}} = 6$ images.

4.2. Ablation Study

Baseline and the varaints of our approach. Our approach is developed based on the model proposed in [44], which adopts ResNet-50 as the backbone architecture, utilizes OIM loss for feature embedding, and outputs a 128-dimensional single-scale feature vector for each image. Based on the model, we build additional six variants of our approach for ablation studies.

Model	Training Loss		scale
	Identifi.	Verifi.	num.
1. basel.	Y	None	1
2. basel.(S)+local	Y	single-scale local sim.	1
3. basel.(M)+local	Y	$(t^1_{p,k} + t^2_{p,k} + t^3_{p,k})/3$	3
4. basel.(M)+group*	N	group sim. $\hat{y}_{p,k}$	3
5. basel.(S)+group	Y	single-scale group sim.	1
6. basel.(M)+local#	Y	group training batches	3
7. basel.(M)+group	Y	group sim. $\hat{y}_{p,k}$	3

Table 1: Detailed configurations for the baseline and other variants.

The configurations of the baseline and the variants are displayed in Table 1. Among them, *basel.* only adopts the identification loss in Eq. (12), *basel.(M)+group** only adopts the verification loss in Eq. (11), and other models employ both identification loss and verification loss. For the methods that adopt verification loss, *basel.(S)+local* and *basel.(M)+local* directly apply the verification loss to supervise the local similarities, while the others apply the verification loss to the group similarities, which indirectly influence the learning of local similarities and feature embeddings. In addition, the models denoted by "(S)" utilize singe-scale feature embeddings while the models denoted by "(M)" employ three-scale feature embeddings.

Our method depends on a special training batch (Sec. 3.6) to construct image pairs for group similarity. To highlight the characteristic of group similarity, the variants employing local similarities utilize randomly shuffled image pairs to compose training batches as previous methods. We also design *basel.(M)+local#* to adopt the same way to construct the image pairs as the proposed method. Results in Table 2m show that data organization is critical, which implicitly connects a group of images for similarity learning and can better discriminate the intra-person variations from inter-person ones in one batch.

Feature embeddings. To investigate how group similarity can benefit the learning of feature embeddings, we compare *basel.*, *basel.(S)+local* and *basel.(S)+group*. For fair comparison, all the methods use single-scale feature embedding $\phi^1(I_i)$ obtained from the MFE module (see Fig. 3) and adopt the Euclidean distance to measure the similarity between two feature embeddings. The feature embeddings of *basel.(S)+local* (Table 2b) consistently improve those of *basel.* (Table 2a), and the feature embeddings of *basel.(S)+group* further improve those of *basel.(S)+local*, where the mAP is increased by 6.9% and 5.2% on the Market-1501 dataset and the DukeMTMC-reID dataset. We employ t-SNE to visualize the feature embeddings of the same 40 testing persons yielded by *basel.(S)+local* and *basel.(S)+group* in Fig. 5, which clearly shows that incorporating the group similarities for training can generate more discriminative feature embeddings.

Local similarities. To investigate whether learning with the group similarity can improve the quality of local simi-

Models	Similarity Metric	Embeddings	Used Modules		Market-1501			DukeMTMC-reID		
			LS	GS	mAP	top-1	top-5	mAP	top-1	top-5
a. basel.	$\|\phi^1(I_p)-\phi^1(I_k)\|_2^2$	$\phi^1(I_i)$	N	N	63.4	83.2	93.8	55.1	74.2	86.2
b. basel.(S)+local	$\|\phi^1(I_p)-\phi^1(I_k)\|_2^2$	$\phi^1(I_i)$	Y	N	70.6	88.2	95.8	59.6	77.2	88.8
c. basel.(S)+group	$\|\phi^1(I_p)-\phi^1(I_k)\|_2^2$	$\phi^1(I_i)$	Y	Y	77.5	90.4	97.2	64.8	80.9	90.8
d. basel.(S)+local	$t_{p,k}^1$	$\phi^1(I_i)$	Y	N	71.1	88.3	96.0	61.9	79.5	89.2
e. basel.(S)+local	$t_{p,k}^2$	$\phi^2(I_i)$	Y	N	70.9	87.7	95.7	59.0	76.8	88.2
f. basel.(S)+local	$t_{p,k}^3$	$\phi^3(I_i)$	Y	N	69.5	86.6	95.8	59.4	77.3	89.1
g. basel.(S)+group	$t_{p,k}^1$	$\phi^1(I_i)$	Y	Y	78.7	91.8	97.2	66.4	81.7	91.0
h. basel.(S)+group	$t_{p,k}^2$	$\phi^2(I_i)$	Y	Y	77.9	91.4	97.1	65.1	80.6	90.8
i. basel.(S)+group	$t_{p,k}^3$	$\phi^3(I_i)$	Y	Y	77.2	91.1	97.2	64.7	80.4	90.3
j. basel.(M)+local	$(t_{p,k}^1+t_{p,k}^2+t_{p,k}^3)/3$	$\{\phi^s(I_i)\}_{s=1}^3$	Y	N	73.8	89.8	96.5	62.9	78.9	90.4
k. basel.(M)+group	$(t_{p,k}^1+t_{p,k}^2+t_{p,k}^3)/3$	$\{\phi^s(I_i)\}_{s=1}^3$	Y	Y	80.5	92.7	97.4	68.0	83.1	91.5
l. basel.(M)+group*	$\hat{x}_{p,k}$	$\{\phi^s(I_i)\}_{s=1}^3$	Y	Y	73.7	86.9	94.9	63.7	79.7	90.1
m. basel.(M)+local$^{\#}$	$(t_{p,k}^1+t_{p,k}^2+t_{p,k}^3)/3$	$\{\phi^s(I_i)\}_{s=1}^3$	Y	N	78.4	92.2	97.6	67.8	82.0	91.9
n. basel.(M)+group	$\hat{x}_{p,k}$	$\{\phi^s(I_i)\}_{s=1}^3$	Y	Y	**81.6**	**93.5**	**97.7**	**69.5**	**84.9**	**92.3**

Table 2: Evaluation of our baseline and its variants on the Market-1501 dataset and the DukeMTMC-reID dataset. We study the influence of multi-scale feature embeddings, different similarity metrics, and training with group similarities. Top-1,-5 accuracies (%) and mAP (%) are reported.

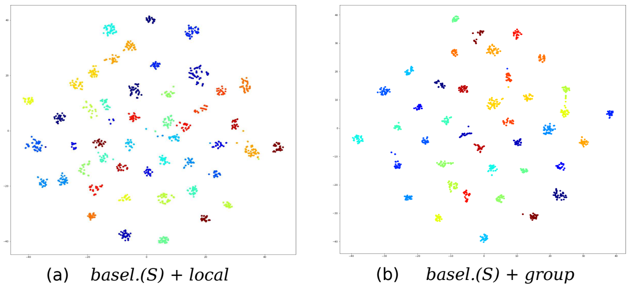

(a) *basel.(S) + local* (b) *basel.(S) + group*

Figure 5: t-SNE visualization of feature embeddings. Each point indicates a testing image from randomly selected 40 identities of Market-1501, and its color indicates the identity. Different identities may share the same color.

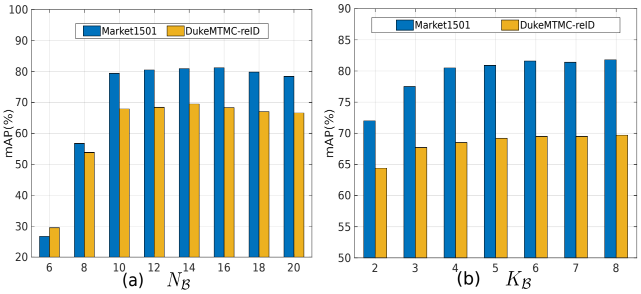


Figure 6: Parameter analysis for group composition. (a) mAP changes with the per-batch person number $N_{\mathcal{B}}$. (b) mAP changes with the per-person image number $K_{\mathcal{B}}$.

larities, we compare *basel.(S)+group* with *basel.(S)+local*, and evaluate them with the feature embeddings of all the three-scale obtained from the MFE module. The main difference between the two modules is that *basel.(S)+local* only consider local constraints for similarity learning, while *basel.(S)+group* depends on the whole group, which indirectly influences the learning of local similarities. The results reported in Table 2 show that *basel.(S)+group* (Table 2 g,h,i) can consistently improve the *basel.(S)+local* (Table 2 d,e,f) of different scales, where the average gain of top-1 accuracy and mAP are 3.9%, 3.0% on the Market-1501 dataset, and 7.4%, 5.3% on the DukeMTMC-reID dataset.

Multi-scale combination. As CRF excels in exploiting diverse information, we utilize the learned coefficients $\{\alpha^s\}_{s=1}^3$ to linearly combine $t_{p,k}^1$, $t_{p,k}^2$ and $t_{p,k}^3$ to obtain $\hat{x}_{i,j}$ (Eq. (10)) for the final similarity between I_p and I_k. We evaluate the proposed weighted combination by comparing it with two different fusion methods: (i) *basel.(M)+local* with the average of local similarities as the similarity metric (Table 2j), (ii) *basel.(M)+group* also with the average of local similarities as the similarity metric (Table 2k). By comparing the variants using multi-scale feature embeddings with the variants using single-scale feature embeddings (Table 2j vs. Table 2 d,e,f, and Table 2k vs. Table 2g,h,i), we observe that the models with multi-scale feature embeddings can generally bring improvements over the models of a single-scale. Among the models with multi-scale feature embeddings, *basel.(M)+group* shows the advantages against *basel.(M)+local* by employing group similarities for training. Besides, CRF learned coefficients lead to better combination (Table 2n) than simply averaging the local similarities of different scales (Table 2k).

Effectiveness of applying identification loss. To evaluate the necessity of employing identification (OIM) loss over the feature embeddings, we construct *basel.(M)+group** by removing the identification loss (Eq. (12)) in the training stage. The gap between the results in Table 2l and Table 2n indicates the identification loss is indispensable in the training stage, which also influences the quality of local similarities and their combination.

Influence of the group composition. In each batch, the person number $N_{\mathcal{B}}$ and the per-person image number $K_{\mathcal{B}}$ together determine the composition of an image group. To study how the composition of the group influences the performance, we first show the mAP changes with $N_{\mathcal{B}}$ by fix-

	Methods	Market-1501		DukeMTMC	
		mAP	top-1	mAP	top-1
H	BoW [49](ICCV15)	14.8	35.8	12.2	25.1
	LOMO+XQDA [23] (CVPR15)	22.2	43.8	17.0	30.8
	SCSP [3] (CVPR16)	26.4	51.9	- -	- -
	DNS [46] (CVPR16)	35.7	61.0	- -	- -
D	Verif.+Identif. [51] (Arxiv16)	59.9	79.5	49.3	68.9
	DCAF [19] (CVPR17)	57.5	80.3	- -	- -
	P2S [55] (CVPR17)	44.3	70.7	- -	- -
	OIM [44] (CVPR17)	- -	82.1	- -	68.1
	GAN [52] (ICCV17)	66.1	84.0	47.1	67.7
	DLPAR [48] (ICCV17)	63.4	81.0	- -	- -
	SVDNet [37] (ICCV17)	62.1	82.3	56.8	76.7
	TriNet [14] (Arxiv17)	69.1	84.9	- -	- -
	JLML [21] (IJCAI17)	65.5	85.1	- -	- -
	SVDNet+REDA [54] (Arxiv17)	71.3	87.1	62.4	79.3
	DPFL [8] (ICCVW17)	73.1	88.9	60.6	79.2
	Proposed approach	**81.6**	**93.5**	**69.5**	**84.9**

Table 3: Comparison with state-of-the-art methods on the Market-1501 and DukeMTMC-reID datasets, which are separated into handcrafted feature based methods (H) and deep learning based methods (D). Top-1 accuracies (%) and mAP (%) are reported.

	Methods	Labelled		Detected	
		top-1	top-5	top-1	top-5
H	BoW [49] (ICCV2015)	18.9	36.2	- -	- -
	LOMO+XQDA [23] (CVPR15)	52.2	- -	46.3	- -
	GOG [28] (CVPR16)	67.3	91.0	65.5	88.4
	DNS [46] (CVPR16)	62.6	90.1	54.7	84.8
	SSSVM [47] (CVPR16)	57.0	84.8	51.2	81.5
D	IDLA [1] (CVPR15)	54.7	86.4	45.0	76.0
	Deep Metric [34] (ECCV16)	61.3	88.5	52.1	84.0
	Gated-SCNN [38] (ECCV16)	- -	- -	68.1	88.1
	DCAF [19] (CVPR17)	74.2	94.3	68.0	91.0
	OIM [44] (CVPR17)	77.7	- -	- -	- -
	CAN [26] (TIP17)	77.6	95.2	69.2	88.5
	JLML [21] (IJCAI17)	83.2	98.0	80.6	96.9
	SVDNet [37] (ICCV17)	- -	- -	81.8	95.2
	DLPAR [48] (ICCV17)	85.4	97.6	81.6	**97.3**
	DPFL [8] (ICCVW17)	86.7	- -	82.0	- -
	Proposed approach	**90.2**	**98.5**	**88.8**	**97.2**

Table 4: Comparison with state-of-the-art methods on the CUHK03 dataset, which are separated into handcrafted feature based methods (H) and deep learning based methods (D). Top-1 and Top-5 accuracies (%) are reported.

ing $K_{\mathcal{B}} = 6$ in Fig. 6a, where small $N_{\mathcal{B}}$ leads inferior results. It is reasonable as too few persons in the group cannot provide sufficient and diverse pairwise relations for the CRF to exploit, making our model hard to train. Besides, we observe that incorporating too many persons in the group also slightly decreases the performance. We also show the influence of $K_{\mathcal{B}}$ by fixing $N_{\mathcal{B}} = 15$ in Fig. 6b, where the overall performance is relatively robust to image number. Even with $K_{\mathcal{B}} = 2$, our approach can still generate satisfactory results. The mAP grows as $K_{\mathcal{B}}$ increases, but maintains stable when $K_{\mathcal{B}} \geq 5$.

4.3. Comparison with State-of-the-art Approaches

We compare the proposed approach with state-of-the-art approaches. The presented results are **not** refined by any post-processing technique such as re-ranking [53] or multi-query fusion [49].

Market-1501 and DukeMTMC-reID. In Table 3, we compare the proposed method with state-of-the-art approaches on Market-1501 and DukeMTMC-reID. It can be seen that deep learning approaches significantly outperform the traditional ones with handcrafted features, while our method further improves the current deep learning approaches by a considerable margin. The compared method DPFL [8] employs multi-scale feature embeddings, whose performance is close to our simplified variant *basel.(M)+local* (Table 2j). Thus the main gains, which have 8.5% and 8.9% mAP on the Market-1501 dataset and the DukeMTMC-reID dataset, are benefited from the employment of group similarity during the training stage.

CUHK03. There are two types of person bounding boxes: one type is manually labeled and the other one is obtained by a pedestrian detector. We report the top-1 and top-5 accuracies in Table 4. Our approach significantly outperforms the compared methods, especially in top-1 accuracy. It is noteworthy that the gap between the labeled evaluation and the detected evaluation of our method is relatively smaller than those of other methods, which indicating that our method is more resistant to the misalignment of bounding box. Besides, DLPAR [48] is 0.1% better than ours on the top-5 accuracy for the detected bounding boxes. One possible reason is that DLPAR adopts the part extractor that is robust to misalignment. It is valuable to combine our approach with such pose-aligned representation for more accurate estimation in the future.

5. Conclusion

We proposed a novel similarity learning approach for person re-identification by combining the CRF model with deep neural networks. The proposed method models relations between images in the group via a unified graphical model, and learns multi-scale local similarities with the aid of group similarities. As more inter-image relations are considered in our model, the learned similarity metric is robust and consistent with images of much variations. Our ablation studies show that our method can learn better feature embeddings, local similarities and multi-scale combination. The proposed method achieves state-of-the-art performance on three public person Re-ID datasets.

Acknowledgement. This work is supported by SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong (Nos. CUHK14213616, CUHK14206114, CUHK14205615, CUHK14203015, CUHK14239816, CUHK419412, CUHK14207814, CUHK14208417, CUHK14202217), the Hong Kong Innovation and Technology Support Program (No.ITS/121/15FX), and China Postdoctoral Science Foundation (Nos. 2017M610641,2014M552339).

References

[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015. 1, 2, 8

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 4

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, 2016. 2, 8

[4] D. Chen, Z. Yuan, G. Hua, N. Zheng, and J. Wang. Similarity learning on an explicit polynomial kernel feature map for person re-identification. In *CVPR*, 2015. 2

[5] D. Chen, Z. Yuan, J. Wang, B. Chen, G. Hua, and N. Zheng. Exemplar-guided similarity learning on polynomial kernel feature map for person re-identification. *IJCV*, 123(3):392–414, 2017. 2

[6] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 40(4):834–848, 2018. 2

[7] W. Chen, X. Chen, J. Zhang, and K. Huang. Beyond triplet loss: A deep quadruplet network for person re-identification. In *CVPR*, 2017. 1, 2

[8] Y. Chen, X. Zhu, and S. Gong. Person re-identification by deep learning multi-scale representations. In *ICCV*, 2017. 8

[9] 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 *CVPR*, 2016. 1, 2

[10] E. Corvee, F. Bremond, M. Thonnat, et al. Person re-identification using spatial covariance regions of human body parts. In *AVSS*, 2010. 2

[11] S. Ding, L. Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993 – 3003, 2015. 2

[12] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 2

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[14] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 1, 2, 8

[15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 4

[16] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 2

[17] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 2

[18] J. Lafferty, A. McCallum, and F. C. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001. 2

[19] D. Li, X. Chen, Z. Zhang, and K. Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *CVPR*, 2017. 2, 8

[20] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014. 2, 6

[21] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017. 2, 8

[22] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith. Learning locally-adaptive decision functions for person verification. In *CVPR*, 2013. 2

[23] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, 2015. 8

[24] G. Lin, C. Shen, A. van den Hengel, and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016. 2

[25] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[26] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *TIP*, 26(7):3492–3506, 2017. 8

[27] B. Ma, Y. Su, and F. Jurie. Bicov: a novel image representation for person re-identification and face verification. In *BMVC*, 2012. 2

[28] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato. Hierarchical gaussian descriptor for person re-identification. In *CVPR*, 2016. 8

[29] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012. 2

[30] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, 2016. 2

[31] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV Workshop on Benchmarking Multi-Target Tracking*, 2016. 6

[32] C. Russell, P. Kohli, P. H. Torr, et al. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009. 2

[33] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 2

[34] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li. Embedding deep metric for person re-identification: A study against large variations. In *ECCV*, 2016. 8

[35] H. O. Song, S. Jegelka, V. Rathod, and K. Murphy. Deep metric learning via facility location. In *CVPR*, 2017. 2

[36] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014. 2

[37] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 8

[38] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, 2016. 1, 2, 8

[39] R. R. Varior, B. Shuai, J. Lu, D. Xu, and G. Wang. A siamese long short-term memory architecture for human re-identification. In *ECCV*, 2016. 2

[40] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang. Joint learning of single-image and cross-image representations for person re-identification. In *CVPR*, 2016. 1, 2

[41] P. Wang, X. Shen, B. Russell, S. Cohen, B. Price, and A. L. Yuille. Surge: Surface regularized geometry estimation from a single image. In *NIPS*. 2016. 2

[42] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu. Shape and appearance context modeling. In *ICCV*, 2007. 2

[43] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl. Sampling matters in deep embedding learning. In *ICCV*, 2017. 2

[44] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang. Joint detection and identification feature learning for person search. In *CVPR*, 2017. 2, 5, 6, 8

[45] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, 2017. 2

[46] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, 2016. 8

[47] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *CVPR*, 2016. 8

[48] L. Zhao, X. Li, Y. Zhuang, and J. Wang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 1, 2, 8

[49] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 6, 8

[50] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 2

[51] Z. Zheng, L. Zheng, and Y. Yang. A discriminatively learned cnn embedding for person reidentification. *TOMM*, 14(1):13, 2017. 2, 8

[52] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 6, 8

[53] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *CVPR*, 2017. 8

[54] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017. 6, 8

[55] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng. Point to set similarity based deep feature learning for person re-identification. In *CVPR*, 2017. 8