Pattern Recognition Letters
journal homepage: www.elsevier.com

Feature Mask Network for Person Re-identification

Guodong Ding[a,**], Salman Khan[b], Zhenmin Tang[a], Fatih Porikli[b]

[a] *Nanjing University of Science and Technology, Nanjing 210094, China*
[b] *Australian National University, ACT 2601, Australia*

ABSTRACT

Person re-identification aims at establishing the identity of a pedestrian from a gallery that contains images of people obtained from a multi-camera system, which has many applications in video surveillance for public security and safety. Many challenges such as occlusions, drastic lighting and pose variations across the camera views, and noise make this task highly challenging. While most approaches focus on learning features and metrics to derive better representations, we hypothesize that both local and global contextual cues are crucial for an accurate identity matching. To this end, we propose a Feature Mask Network (FMN) that takes advantage of ResNet high-level features to predict a feature map mask and then imposes it on the low-level features to dynamically re-weight different object parts for a complementary feature representation. This serves as an attention mechanism by allowing the network to focus on local details selectively. We frame the network training as a multi-task objective optimization, which further improves the learned feature descriptions. We conduct experiments on Market-1501, DukeMTMC-reID and CUHK03 datasets, where the proposed approach respectively achieves significant improvements and competitive results when compared to the state-of-the-art.

Keywords: person re-identification, image retrieval, network ensemble

1. Introduction

Person re-identification deals with the task of matching identities across images captured from disjoint camera views. Given a query image, a person re-identification system determines whether the person has been observed by another camera at another time. Significant attention has been dedicated to person re-identification in the past few years [3, 57, 51, 6, 27, 7], as the task is essential in video surveillance for cross-camera tracking, multi-camera event detection, and pedestrian retrieval for public security and safety. Albeit highly important, the re-identification problem poses significant challenges due to the viewpoint and pose changes, illumination variations, cluttered backgrounds, occlusions, and indiscriminative appearances across different cameras (or even for the same camera). Moreover, re-identification task considers new identities at test time, therefore it requires a high generalization capability of the learned feature encodings.

Existing research on person re-identification problem can be broadly divided into two mainstreams. **(a)** Methods that seek to learn a discriminative metric, which allows instances of the same identity to be closer while instances of different identities to be far away [5, 12, 18, 49, 44, 16]. These metric learning methods mainly adopt pairwise [12, 18, 49] or triplet [44, 16, 35] loss to obtain an embedding for each probe image and distinguish identities in the projected space. Along similar lines, [3] proposes a quadruplet ranking loss that is capable of achieving smaller intra-class variations and large inter-class distances, which results in an improved performance on the test set. **(b)** Methods that focus on designing robust visual descriptors to model the appearance of the person [31, 8, 52, 25]. Among these techniques, handcrafted features found ini-

[**] Corresponding author.
e-mail: guodong.ding@njust.edu.cn (Guodong Ding)

tial success [8, 31]. More recently, automatically learned feature representation using deep architectures have shown excellent improvements [55, 39, 47, 58]. The prevalent re-identification approaches belonging to these two research streams assume that the person bounding boxes are provided by a dedicated detector. However, such detections are not always perfect, resulting in problems such as the inclusion of excessive background in the object box, incomplete coverage of body and localization mismatch. This is exacerbated by the fact that there exist heavy occlusions partially masking the pedestrians in surveillance scenarios.

Our intuition is that a desired capability that allows overcoming these challenges is to pay attention to important yet perhaps subtle local details alongside the supposedly prominent global cues. In this paper, we propose an automatic approach which learns to focus on complementary local details as well as the global image descriptions using deep neural networks. This helps the identification algorithm to excavate more of the regions that carry more valuable and discriminative cues for the identity prediction task.

We formulate the proposed feature selection strategy as a soft-attention with in a deep network, which enables an end-to-end learning framework. In addition to avoiding the problems due to imperfect pedestrian detection windows, our network learns to resolve ambiguities (such as similar clothing of two different identities) by latently shifting attention towards more distinguishing aspects of the respective identities. To this end, we utilize already learned global discriminative features as a guidance and a dynamic selection mechanism to assign different importance weights to low-level features for a more compact human feature representation.

Network Ensemble has also been proposed as a technique where multiple classifiers are combined for a better representation learning [13, 19]. The diversity amongst these classifiers is playing a vital role. To this end, we propose a multi-task loss formulation that considers both classification and pairwise ranking objectives during the training phase. The inter-branch pairwise ranking loss enforces the counterpart network outputs to take guidance from the predictions based on global features by introducing a margin violation penalty to promote diversity between these branches. Our results on three large-scale datasets demonstrate significant performance improvements. Our main contributions are threefold, as given below:

- We propose a Feature Mask Network (FMN) that can dynamically attend to complementary local details in an image and use them alongside global representations for improved pedestrian re-identification.
- We introduce a multi-task formulation, which optimizes a classification as well as a single input based inter-branch pairwise ranking loss to learn highly robust feature descriptions.
- The proposed approach is easy to implement, efficient to train, while achieving comparable results compared to the state-of-art methods on all three benchmark datasets.

The remainder of this paper is organized as follows. A review of related work is provided in Section. 2. Section 3 presents our proposed feature mask network approach. Section 4 exihibits the effectiveness of proposed method and provides an extensive ablation study. We conclude our work in Section 5.

2. Related Work

In the past few years, many efforts have been proposed for the task of person re-identification, which greatly advance this field. The main aspect of this force is discriminative feature representation learning [58, 44, 39, 47, 4]. Existing approaches seek to design a robust human descriptor against illumination and viewpoint variations as well as misalignments. In light of stronger feature representation learning, we next review two closely related categories in the following parts.

Fine-grained re-id learning: Person re-identification task is difficult in the fact that a robust pedestrian descriptor is hard to obtain given one identity's varying poses and drastic viewpoint changes. Recently, attention mechanism has been adopted by several works [45, 24, 52, 20, 30, 21, 38] to address this problem. Shifting focus on different regions of a person image can help the descriptor gain more discriminate ability. The common practice of these methods is to select attentive regions which is beneficial to the re-id task. For example, Wei et al. [45] first extract coarse human regions based on body key points and then perform region specific feature learning along with the full image branch to obtain more discriminative representations. Li et al. [21] adopt similar approach locating latent discriminative regions and exploit these regional features for performing re-id. While Su et al. [38] selects re-id discriminative parts using Spatial Transformer Network (STN) [17] following a hard attention manner. While our work generates low-level feature re-weighting parameters from high-level id discriminative feature directly which can bypass the rigid hard region selection thus attain soft and flexible attentive regions.

Network Ensemble: Neural network ensemble is a technique where multiple models are created and combined to obtain an improved representation, compared to creating just one model [32]. Existing research, both theoretical [13, 19] and empirical [14, 33], has shown that an ensemble is generally more accurate than any of the individual classifiers within the ensemble. For example, Sollich and Krogh [37] showed ensembles tend to yield better results when there is a significant diversity among the models. *Therefore, a good ensemble seeks to promote diversity among the models they combine.* Recently, Zheng et al. [58] proposed a Pedestrian Alignment Network exploiting STN [17] to apply transformation on a second branch feature maps to deal with the misalignment problem, and then combine features from both branches as the final descriptor. In essence, their approach can be regarded as an instance of network ensembling. Our approach shares the same spirit but differs in the consideration of channeling both branches along with a self-learned mask for latent feature selection within the network.

3. Proposed Method

Our approach is based on the proposition that a successful person re-identification system needs to give importance to both the global and the complementary local discriminative aspects of a pedestrian whose image is acquired from different views using multiple surveillance cameras. To this end, we introduce a novel CNN based deep learning architecture, which learns to focus both on the global and complementary cues of a person that are useful for its re-identification. We describe our proposed architecture and training procedure in the following sections.

3.1. Feature Mask Network

The proposed CNN architecture is shown in Fig. 1. The complete system consist of four main components, which we term as **a)** Base Network, **b)** Global Representation Network, **c)** Mixing Network and **d)** Counterpart Representation Network.

The Base Network (BN) is formed by a convolution and a max pooling layer to learn rich low-level image representation as inputs to subsequent computational blocks within our network.

The Global Representation Network (GRN) learns the holistic feature representations corresponding to an input image. It is designed as a Residual Network [15] with five residual blocks each (except the first module which is used as BN) containing skip connections. The GRN has a total of 48 parameter layers with a total of 3.8 billion FLOPs. It has been pre-trained on the ImageNet image classification dataset and fine-tuned on the person re-identification dataset.

The Mixing Network (MN) predicts the mask weights for the local features from the initial layers in the GRN. These weights are derived from the global feature representation learned using the GRN for the pedestrian images. The MN consists of a transformation layer implemented as a fully connected (FC) layer on top of the global features from GRN, followed by a reshaping module and a mixer which performs element-wise product between the counterpart feature and the mask weights.

It has been studied that a network tends to focus on the most discriminative parts of an image when performing classification [36]. As a result, holistic representations outputted from GRN mainly focus on human torso. While input pedestrian images may contain excessive background or misalignment errors due to high appearance, scale and pose variations in the candidate profiles. Therefore, the important information in an image may get suppressed. The third block in our scheme, called the Counterpart Representation Network (CRN), learns to attend to local discriminative features which can provide useful clues for a person's identity matching. A CRN takes output features from BN, which are re-weighted by the mask predicted using the MN. These modified activations basically exhibit an attention mechanism, where useful discriminating counterpart features are given more importance compared to others. The CRN consists of four residual block with identical architecture to the corresponding blocks in the GRN. Since, the CRN parameters are not shared with the GRN, it learns a *different global representation with a refocus on the locally discriminating information.* This information acts as a complementary source of information which we found to be highly useful in our experiments (see Sec. 4.4 for discussion). In the following, we describe the details of mask computation.

3.2. Mask Computation

The MN operates on the global feature representation $\mathbf{g} \in \mathbb{R}^m$ from higher layers (final fully connected layer in our case) and the low-level feature representation $\mathbf{f} \in \mathbb{R}^n$ from Base Network. Since the feature representations from the lower layers in a CNN is arranged as multiple 2D activation maps for color images, we can represent their dimensions more conveniently as: $n = h \times w \times c$, where h, w and c denote the height, width and the number of feature channels respectively. The MN first transforms the global feature from GRN to a $n' = h \times w$ dimensional output which can then be used to compute the feature mask:

$$\mathbf{m'} = \sigma(\mathbf{W}^{\mathrm{T}}\mathbf{g}), \tag{1}$$

where, σ denotes the ReLU activation function and $\mathbf{m'} \in \mathbb{R}^{n'}$ is the transitional mask and $\mathbf{W} \in \mathbb{R}^{m \times n'}$ is the weight matrix for the transformation which is equivalent to a fully connected layer in our mixing network. As a result, we can learn this feature mapping directly from the data which can provide an *image-specific* mask for the counterpart feature representation learning. Since our goal is to attend to extra local features in the spatial domain, we identically re-weight all feature channels in $\mathbf{f}$ using the same predicted feature mask. The final feature mask ($\mathbf{m} \in \mathbb{R}^{h \times w}$) is obtained from $\mathbf{m'}$ by reshaping and applying element-wise exponentiation as follows:

$$\begin{aligned} &\mathbf{m}_{i,j} = \exp(\mathbf{m}_k), \\ &s.t.,\ j = \lfloor k/h \rfloor + 1,\ i = k - h(j-1) \end{aligned} \tag{2}$$

Once the feature mask $\mathbf{m}$ is obtained, MN uses a mixer to combine it with $\mathbf{f}$ using the channel-wise Hadamard product (denoted as '$\circ$' below):

$$\mathbf{o}_i = \mathbf{f}_i \circ \mathbf{m}, \quad s.t.,\ i \in [1, c] \tag{3}$$

where, i denotes the feature channel number in $\mathbf{f}$. We outline the training process for our proposed network below.

The role of mask: We believe mask plays two main roles in our proposed architecture, 1) it feeds already obtained global information embedding $\mathbf{g}$ into counterpart

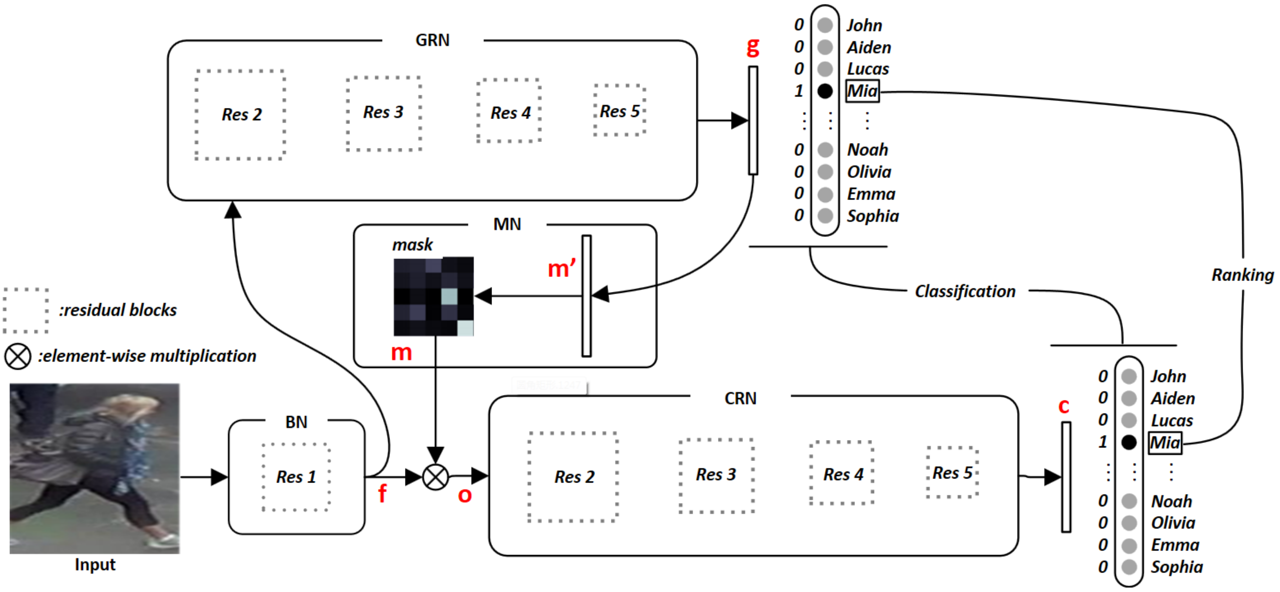

Fig. 1. The overall architecture of our network. Residual blocks in dashed box correspond to blocks defined in a ResNet-50 architecture. The GRN obtains high-level feature representations. Using these features, the MN creates a feature mask which dynamically attends to complementary details useful to person identification. The re-weighted features are used in the bottom CRN to learn more locally focused features. Note that the GRN and CRN share same architecture yet learn a separate set of parameters. During testing, we concatenate the last fully connected layer features from both branches to form the final pedestrian descriptor.

branch, thus counterpart branch can learn extra supplementary representation, enabling the diversity of components in a network ensemble; 2) the mask is imposed spatially, which serves as a latent spatial attention mechanism to enable CRN the ability to dynamically attend to different parts. This is intuitive because identity specific details are not clustered in a single image region, but distributed along the human body e.g., footwear details and apparels.

3.3. Classification and Ranking

The network is trained in two stages, summarized in Algorithm 1. *First*, the GRN is trained to predict the pedestrian identities in an end-to-end manner. The parameter learning process for GRN involves a weight initialization step using a ResNet-50 model pretrained on the ImageNet dataset, followed by our task-specific fine-tuning. Once the training process is complete, the feature representation from the GRN encodes global discriminator information corresponding to a given image. *Afterwards*, the GRN weights are kept fixed, while the MN and CRN weights are learned jointly in the next stage. Similar to the first stage, the second stage training is also performed using the pedestrian identities as the ground-truth labels. In contrast to the global representations learned via GRN, the second stage training is desigened to discover more complementary discriminative information and re-shifts the attention appropriately using the MN to obtain a complementary feature representation.

Both the GRN and CRN are trained for the identity classification task using supervised learning in two stages. We use a conventional cross-entropy loss function for both stages as follows:

$$\mathcal{L}_{cls}(\mathbf{p},\mathbf{y}) = -\sum_{k} y_k \log \frac{\exp(p_k)}{\sum_j \exp(p_j)}, \qquad k \in [1,r] \quad (4)$$

where $\mathbf{p} \in \mathbb{R}^r$ denote the predicted output and $\mathbf{y} \in \mathbb{R}^r$ denotes the desired output as a one-hot vector. Here, r denotes the number of classes in the dataset, which is equal to the total units in the output layer.

To better utilize GRN as well as promote diversity between GRN and CRN, in the second stage of joint network training, we propsoe to add on top a inter-branch pairwise ranking loss defined as follows :

$$\mathcal{L}_{rank}(p_t^G, p_t^C) = \max(0, p_t^G - p_t^C + m), \qquad (5)$$

where m represents the margin, p_t^G, p_t^C denote the prediction probability on the correct category label t of GRN and CRN, respectively. Imposing rank loss enables CRN to take the prediction form GRN as reference, and enforces CRN to make better predictions for correct labels by a margin, thus leading to more confident and accurate predictions.

3.4. Image Descriptor

After learning the parameters of the network, the final image image descriptor at test time is obtained by combining the feature representations from the GRN and the CRN. These feature representations are derived from the last FC layer in each network, which contain task specific discriminative information pertaining to both global and local pedestrian attributes. The following relation is used

Algorithm 1: Feature Mask Net Optimization

```
Input: Pre-trained model $\phi(\cdot;\boldsymbol{\theta_o})$, Re-ID training
       data $I$, Identity labels $Y$, Maximum
       Iterations $T_g$, $T_c$ for GRN and CRN,
       respectively.
Output: Learnt FMN Model $\phi(\cdot;\hat{\boldsymbol{\theta_b}},\hat{\boldsymbol{\theta_g}},\hat{\boldsymbol{\theta_m}},\hat{\boldsymbol{\theta_c}})$, $\hat{\boldsymbol{\theta_b}}$,
        $\hat{\boldsymbol{\theta_g}}$, $\hat{\boldsymbol{\theta_m}}$ and $\hat{\boldsymbol{\theta_c}}$ denote parameters of BN,
        GRN, MN and CRN, respectively.
Initialization: $\boldsymbol{\theta_o} \rightarrow \boldsymbol{\theta_b}, \boldsymbol{\theta_g}, \boldsymbol{\theta_c}$, random initialization
for $\boldsymbol{\theta_m}$
Global representation learning;
for $t = 1 : T_g$ do
    Keep $\boldsymbol{\theta_m}$, $\boldsymbol{\theta_c}$ fixed;
    Update $\boldsymbol{\theta_b^t}$,$\boldsymbol{\theta_g^t}$ using Eq. (4)
$\hat{\boldsymbol{\theta_b}} \leftarrow \boldsymbol{\theta_b^{T_g}}$,$\hat{\boldsymbol{\theta_g}} \leftarrow \boldsymbol{\theta_g^{T_g}}$ ;
Counterpart representation learning;
for $t = 1 : T_c$ do
    Keep $\hat{\boldsymbol{\theta_b}}, \hat{\boldsymbol{\theta_g}}$ fixed and feed-forward GRN;
    Produce mask $\boldsymbol{m^t}$ using Eq. (1) and Eq. (2);
    Apply mask on CRN inputs with $\boldsymbol{m^t}$ using Eq.
     (3);
    Update $\boldsymbol{\theta_m^t}$ and $\boldsymbol{\theta_c^t}$ using both Eq. (4) and Eq.
     (5);
$\hat{\boldsymbol{\theta_m}} \leftarrow \boldsymbol{\theta_m^{T_c}}$, $\hat{\boldsymbol{\theta_c}} \leftarrow \boldsymbol{\theta_c^{T_c}}$ ;
Return: $\phi(\cdot;\hat{\boldsymbol{\theta_b}},\hat{\boldsymbol{\theta_g}},\hat{\boldsymbol{\theta_m}},\hat{\boldsymbol{\theta_c}})$
```

to compute a re-weighted concatenation of normalized individual descriptors:

$$D = [\alpha \frac{\mathbf{g}^{\mathrm{T}}}{\parallel \mathbf{g} \parallel_2}, (1-\alpha)\frac{\mathbf{c}^{\mathrm{T}}}{\parallel \mathbf{c} \parallel_2}]^{\mathrm{T}}, \tag{6}$$

where $\parallel \cdot \parallel_2$ operator denotes an ℓ^2-norm. The parameter α decides the trade-off between features $\mathbf{g}, \mathbf{c}$ from the two separate breaches, GRN and CRN, respectively. We set α to 0.5 in our experiments following [58]. The resulting descriptor is used to find closest matches from the gallery by performing a nearest neighbour (NN) search based on Euclidean distance.

3.5. Discussion

- **Ranking loss choices.** The ranking loss can be applied in two different ways which differ in terms of inputs to loss. The ***first*** kind of ranking loss widely adopted in person re-identification is commonly imposed on the last fully connected layer activations (e.g., 'pool5' in ResNet-50 case). It requires image triplets as input and is designed to ensure the positive pair to have higher similarity than negative pair. Thus, it seeks smaller intra-class distances than inter-class distances. However, it suffers form dramatic data expansion when constructing sample triplets from the training set [16]. The ***second*** kind of loss, we call inter-branch pairwise ranking, is imposed on softmax predictions of the same image through a Siamese-like network towards fine-grained classification [10]. Such a design enables network to take prediction from global scale as reference, and approach the supplementary discriminative regions by enforcing the local-scale network to generate more confident predictions. This kind of ranking loss is easy to implement, efficient to train compared to pairwise or triplet approaches and results in performance improvements as demonstrated in our experiments.
- **Joint supervision.** The softmax classification loss does not directly consider the ranking errors and ignores the fact that CRN is learning extra supplementary features with the help of global information. Thus the CRN should have more confident predictions for correct categories. Joint supervision better fits with the structure of our network and leverages relations between two branches. Remarkably, the imposition of ranking loss does not mean that local features should be more important than global features. Instead, these features are assigned a tradeoff parameter to balance their contributions when concatenated to form the final image descriptor.

4. Experiments

4.1. Datasets

Market-1501: Market-1501 is a person re-identification dataset consisting of 32,688 images with bounding boxes of 1501 pedestrians. These image are captured on campus and persons are detected using Deformable Parts Model (DPM) [9]. In this experiment setting, 12,936 detected images of 751 identities been taken as training set, whilst the rest 19,732 images of 750 identities are used for testing, following the original evaluation protocols.

DukeMTMC-reID: DukeMTMC-reID is one of the largest pedestrian image datasets derived from DukeMTMC [34] and comprises of 36,411 pedestrian images of 1,404 identities. It is split into train/test sets of 16,522/19,889, and we evaluate with 2,228 queries to retrieve from 17,661 gallery images. The dataset introduces occlusion of pedestrians, such as cars or trees, which makes this task more challenging.

CUHK03: CUHK03 contains 14,097 images of 1,467 identities. On average, each identity has 9.6 images captured by 5 different sets of cameras. The dataset provides both manually cropped bounding boxes and the automatically cropped bounding boxes using a pedestrian detector, named as '*CUHK03 labeled*' and '*CUHK03 detected*', respectively. For fair comparison, we evaluate our method on the new training/testing protocol proposed in [59] which divides original dataset by half, yielding a training set of 767 identities and a testing set of 700 identities. [59] argues that this split conforms with the real-world person re-id tasks, which can only provide limited training samples, while re-id is performed on a larger unseen sample pool.

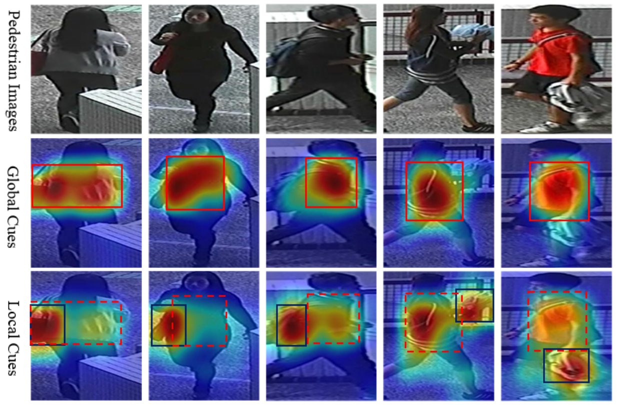

Fig. 2. While global information such as apparel and body shape can be useful for re-identification in some cases, local details such as backpack, handbag, clothing-style and shoes can be even more helpful in other cases. Best viewed in color.

Table 1. Rank-1 accuracy and mAP results with different BN structure selected on Market-1501 dataset.

layer	mask size	rank-1	mAP
Pool1	56×56	**86.00**	**67.12**
Res2	56×56	85.15	65.84
Res3	28×28	84.23	65.16
Res4	14×14	82.39	61.95

4.2. Implementation Details

Network Training: Due to the high classification accuracy of ResNet-50 [15], following the baseline in [55] we also use it as our backbone architecture. The network is pre-trained on the ImageNet dataset consisting of 1000 object classes. To fine-tune it for the person re-identification task, we replace the 1000 units in the final FC layer with the number of training identities in dataset. Note that in our case, BN together with GRN or CRN composes a whole ResNet-50 architecture, that is to say, our proposed model can be regarded as two ResNet-50 models sharing only Base Network part. BN and GRN branch in our networks are trained with an initial learning rate of 0.1, that is reduced to 0.01 after 20 epochs. For the MN and CRN branches, we train with the same initial learning rate, which is reduced to 0.01 after 35 epochs. The training is performed until the the network converges. We update our parameters with stochastic gradient descent with 0.9 momentum. The training dataset is augmented with horizontal flipping and cropping the original images.

Evaluation Metrics: We evaluate our methods with mean average precision (mAP) and rank-1, rank-5 and rank-20 accuracy measures. The rank-i accuracy denotes the rate at which one or more correctly matched images appear in top-i. If no correctly matched images appear in the top-i of the sorted list, rank-i=0, otherwise rank-i=1. We report the mean rank-i accuracy for query images. Also, for each method, we calculate the area under the Precision-Recall curve and the mean of the average precision scores for each query, which reflects the overall precision and recall rate.

The proposed framework is implemented using the MatConvNet [43] library.

4.3. Ablative Study

In this section we evaluate the effectiveness of each component in our architecture.

Baseline: Our baseline network is adopted and fine-tuned on ResNet-50 by substituting final fully-connected layer from 1000 to target identity numbers, denoted as GRN, to achieve 79.33% , 67.91%, 38.00% and 34.36% respectively for four evaluated datasets. These results denote strong and competitive rank-1 accuracies.

GRN vs CRN: The individual performance of the GRN, CRN results of the model are shown in Table 2. As one can note, independent CRN feature achieves score close to or less than the baseline branch, we speculate this is because CRN mainly focuses on local features. As shown in Fig. 2, red dashed rectangle boxes respectively correspond to features learned by GRN, and it is clear that global cues in varying degrees has faded, newly learned local features denoted by black boxes are receiving higher activations. This confirms our intuition about learning complementary information from both global and local branches, and leads us to form final pedestrian descriptors in the form of concatenation. Notably, the rank-1 CRN accuracy for CUHK03 detected dataset experience a considerable decline (around 6.5%) compared with the baseline, we speculate that this behavior is due to model overfitting on the training set since we used identical experiment hyperparameters for all datasets while CUHK03 has around half the samples of DukeMTMC-reID for training. Even so, CRN still learns extra information supplementary to GRN since FMN achieved a performance boost.

Ranking: The inter-branch pairwise ranking takes into consideration inherent relations between two branches and helps improve the results. Names with superscript $^-$ denote CRN and FMN models only trained with conventional the softmax loss and rest of the proposed models are trained with joint supervision. Overall consistent improvements on four datasets can be observed when joint supervision is applied. However, performance boost is relatively small i.e., around 1-2% improvement in rank1 accuracy. This shows that although a multi-loss formulation is helpful for person re-identification, the improvements due to the combination of complimentary CRN and GRN features are comparatively more significant.

Low-level Feature Selection: A CNN network extracts a hierarchy of features corresponding to different level of details in the input image. From initial to final layers, the output features transition from local to global information. We evaluate the effect of low-level feature output from BN on the person re-identification accuracy. In Table 1, we report rank-1 accuracy and mAP when the local features from different BN structures are used in the mixing network. As aforementioned, BN and GRN form a complete ResNet-50 model, we will refer to last layer of BN

Table 2. Comparison of different methods on Market-1501, DukeMTMC-reID, CUHK03 (detected), and CUHK03 (labeled). Consistent improvement of our proposed on all datasets in terms of rank@k accuracy and mAP can be observed. GRN and GRN$'$ denote same single branch trained individually, NE denotes network ensemble comprising of GRN and GRN$'$, $^-$ denotes network without inter-branch pairwise ranking.

Methods	dim	Market-1501				DukeMTMC-reID				CUHK03 detected				CUHK03 labeled			
		1	5	20	mAP	1	5	20	mAP	1	5	20	mAP	1	5	20	mAP
GRN	2048	79.33	91.48	96.62	58.50	67.91	81.33	89.59	48.40	38.00	48.14	59.93	33.68	34.36	46.64	58.14	30.14
GRN$'$	2048	79.42	90.23	95.88	58.35	68.02	81.46	89.66	49.02	37.17	47.28	58.82	33.71	33.98	46.25	57.89	29.86
CRN$^-$	2048	80.28	90.07	94.71	59.23	68.88	82.05	89.86	50.62	32.71	45.03	59.00	29.69	31.29	47.04	61.27	30.22
CRN	2048	81.15	91.39	96.44	59.88	69.84	83.03	90.04	51.05	31.50	46.07	60.86	29.39	32.36	47.57	61.93	30.41
NE	4096	82.00	92.27	96.34	62.68	71.19	83.98	91.94	53.23	40.33	50.45	62.56	35.74	36.68	49.02	62.58	34.14
FMN$^-$	4096	85.12	93.08	96.57	65.93	73.61	**85.19**	92.06	56.55	40.79	53.00	64.57	37.94	39.55	53.79	65.20	37.66
FMN	4096	**86.00**	**93.74**	**97.51**	**67.12**	**74.51**	85.05	**92.41**	**56.88**	**42.57**	**56.21**	**67.36**	**39.21**	**40.71**	**54.57**	**65.50**	**38.05**

Table 3. Rank-1 accuracy (%) and mAP (%) on Market-1501 dataset. The best and second best performances are marked in red and blue, respectively.

Method	rank-1	mAP	Method	rank-1	mAP
DADM [39]	39.4	19.6	DeepTransfer [11]	83.7	65.5
BoW+kissme [54]	44.42	20.76	GAN [57]	83.97	66.07
MR-CNN [41]	45.58	26.11	PAN [58]	82.81	63.35
MST-CNN [29]	45.10	-	APR [26]	84.29	64.67
FisherNet [46]	48.15	29.94	Triplet[16]	84.92	69.14
CAN [28]	48.24	24.43	PAN+re-rank[58]	85.78	**76.56**
SL [2]	51.90	26.35	JLML [22]	85.1	65.5
DNS [50]	55.43	29.87	DPFL [4]	88.9	73.1
Gate Reid [42]	65.88	39.55	GLAD [45]	**89.9**	73.9
SOMAnet [1]	73.87	47.89	HA-CNN[23]	**91.2**	75.7
PIE [53]	78.65	53.87	Baseline	79.33	58.50
Verif.-Identif. [56]	79.51	59.87	Ours	86.00	67.12
SVDNet [40]	82.3	62.1	Ours+re-rank	87.92	**80.62**

Table 4. Rank-1 accuracy (%) and mAP (%) on DukeMTMC-reID and CUHK03 dataset with new evaluation protocol. New protocol divide each CUHK03 dataset roughly by half for training and testing. Under this setting, we use a larger testing gallery and a smaller training set.

Method	DukeMTMC-reID		CUHK03 Detected		CUHK03 Labeled	
	rank-1	mAP	rank-1	mAP	rank-1	mAP
BoW+kissme [54]	25.13	12.17	-	-	-	-
BoW+XQDA [54]	-	-	6.4	6.4	8.0	7.3
LOMO+XQDA [25]	30.75	17.04	12.8	11.5	14.8	13.6
ResNet+XQDA [59]	-	-	31.1	28.2	32.0	29.6
[59]+re-rank	-	-	34.7	37.4	38.1	40.3
GAN[57]	67.68	47.13	-	-	-	-
OIM[48]	68.1	-	-	-	-	-
APR[26]	70.7	51.9	-	-	-	-
PAN[58]	71.6	51.5	36.3	34.0	36.9	35.0
PAN+re-rank [58]	75.9	**66.7**	41.9	**43.8**	43.9	**45.8**
DPFL [4]	79.2	60.6	40.7	37.0	43.5	40.5
HA-CNN [23]	**80.5**	63.8	41.7	38.6	**44.4**	41.0
Baseline	67.91	48.40	38.2	34	34.4	30.1
Ours	74.51	56.88	**42.6**	39.2	41.0	38.1
Ours+re-rank	**79.52**	**72.79**	**47.5**	**48.5**	**46.0**	**47.6**

in the convention of ResNet-50 (e.g. Pool1, Res1, Pool5). We can see a consistent performance boost as the selected layer (BN output) is changed from final layers towards initial layers. Specifically, the rank-1 accuracy of 82.39% and mAP of 61.95% at Res4 changes to 86.00% and 67.12%, respectively, at Pool1. This observation is intuitive, since the network achieves best performance when the complementary information from both global and local levels are combined for the task. The high layers encode more holistic information about appearance and shape, while the local information can disambiguate cases with high appearance similarities. When local features from higher layers are used, the resulting performance is low since the features contain less details.

Comparison with close work: Closest to our proposed approach is the recent pedestrian alignment network by Zheng et al. [58]. However, the way both techniques learn feature representations are significantly different since they aim to solve the *misalignment problem* in automatically detected pedestrian bounding boxes. This misalignment is reduced by applying an affine transformation learned automatically within a network. In contrast, our work advocates that the intermediate feature representations within the network have different levels of relevance for our end-task since the object of interest often only covers a part of the input image. Therefore, directly excavating complementary features can result in a better performance without the need to align or transform the input image. In our work, we show that appropriately shifting the network's attention towards the local information in feature encoding can greatly help the person re-identification task. Furthermore, the proposed multi-task formulation helps in learning discriminative features by considering the predictions at both global and local levels. In summary, **1)** [58] boils down to a Spatial Transformer Network (STN)[17] variation predicting only six parameters to crop and transform feature maps thus resolving misalignments caused by auto detectors. **2)** Our proposed multi-task formulation is different from [58] , which helps in learning complementary features at global and local levels. **3)** Our approach clearly outperforms [58] on three major person reID datasets with exact same baseline model.

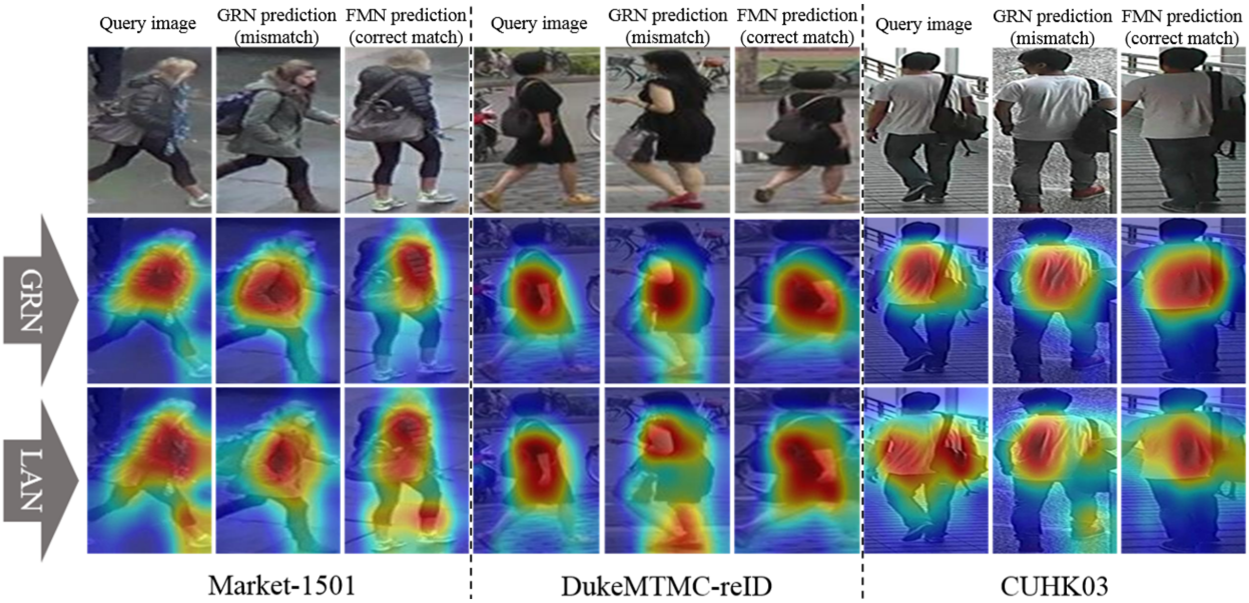
Query image
GRN prediction (mismatch)
FMN prediction (correct match)
Query image
GRN prediction (mismatch)
FMN prediction (correct match)
Query image
GRN prediction (mismatch)
FMN prediction (correct match)
GRN
LAN
Market-1501
DukeMTMC-reID
CUHK03

Fig. 3. Qualitative examples of improved retrieval results on the Market-1501, DukeMTMC-reID and CUHK03 datasets (*left* to *right*). For each query image, we present its false rank-1 match based solely on GRN (denoted by mismatch) followed by its accurate match using our proposed architecture. Second and third rows visualize the heat maps obtained from GRN and CRN, respectively. One can observe that by attending to local distinctive parts (e.g., subtle variations in apparel design, shoes and carry-on bags) using CRN, the overall performance is boosted and the proposed network is therefore better suited for re-identification task.

4.4. Evaluation

Effectiveness of FMN. We evaluate our proposed FMN on all three large-scale re-identification benchmarks to show the effectiveness of our proposed network architecture. When features from both branches are combined (row FMN in Table. 2), a significant boost in overall performance is observed. Rank-1 accuracy for Market-1501, DukeMTMC-reID, CUHK03 detected and CUHK03 labeled datasets have been improved by a margin of 6.66%, 6.60%, 4.57% and 6.35%, respectively. The performance in terms of mAP values have also been improved remarkably by 8.62%, 8.48%, 5.53% and 7.91%, respectively.

NE vs FMN: We regard our proposed FMN as an instance of network ensemble (NE) technique and we provide numerical evidence showing the improvements it gains on three datasets as reported in Table. 2. It is observed that both GRN and GRN$'$ yielded close accuracy and mAP values, which is expected as they have exact same architecture (BN + GRN) but trained individually. Two counterpart representations are later fused using Eq. 6 to form pedestrian descriptor (row NE in Table. 2). Despite the fact these two branches are independent, NE gained a performance boot by an average margin of 2-3% on both evaluation metrics. While with our proposed FMN enabling associations between two branches, the performance is further boosted by another 4.00%, 3.32%, 2.27% and 4.03% on evaluated datasets, respectively, in rank-1 accuracy and overall 4% increase in mAP. This consistent performance gain proves that the counterpart feature representations learned by CRN are complementary to the global features from GRN.

Alongside the quantitative comparisons on the three aforementioned datasets, we also qualitatively analyze the effect of FMN for the case of highly confusing pedestrian examples and show the complementary propriety of learned features. To this end, we illustrate query images from each of the three datasets in Fig. 3 along with the Rank-1 mismatch (predicted by baseline) and a true match (predicted by our proposed model) for the respective pedestrian. We also visualize the heat maps obtained from both the GRN and CRN to study the salient image regions which are given more attention by the network during the prediction process. One can notice from both Fig. 2 and Fig. 3 that global representations from GRN focus mainly around the main torso of the body. This can lead to incorrect predictions because the upper trunk of the body can be identical for two altogether different identities. In contrast, more subtle local details such as the clothing and attire specifics (*left* example), footwear (*middle* example) and differences of the back-packs (*right* example) can provide more useful cues for correct identification of persons. The FMN attends to both global and local details and leverages their complementary characteristics to correctly identify the corresponding match for a query image.

Comparison with the State-of-the-Art Methods: We also compare our proposed approach with the state-of-art methods on Market-1501, DukeMTMC-reID and CUHK03 datasets in Tables 3 and 4 respectively. On all three datasets, we achieve the competitive performances in comparison to the very recent methods with more sophisticated pipelines. As shown in Table 3, we achieved **rank-1 = 87.92%, mAP=80.62%** using the re-ranking approach, which is a highly consistent performance across both metrics. DPFL [4] outperforms us by 1% mainly because: **1)** They train network at multiple scales, which is expected to achieve better accuracy than ours with single scale; **2)** The selection of stronger backbone architecture, our ResNet-50 baseline performance (rank-1 79.3%) is strong and competitive, while their inception-v3 achieved 83.3%, around 4% higher. GLAD [45] outperformed our approach by 2% in rank-1, we speculate this is because it trains with a four-stream network on global and local regions and shared kernels between global and local better help prevent overfitting. On DukeMTMC-reID, our proposed method also achieved the best performance with **rank-1 = 79.52%, mAP = 72.79%**. HA-CNN [23] achieves the highest rank-1 score since a cross-attention (between local and global) interaction learning scheme is introduced for a stronger representation, while our approach adopts a conciser architecture. On CUHK03 dataset, we observe our rank-1 is **5.6%** higher than best competing methods PAN [58] on detected and HA-CNN [23] by **1.6%** on labeled datasets.

5. Conclusion

A person occupies only a portion of the input image, and a global scene description does not suffice for an accurate identity matching. In this work, we proposed a hybrid architecture for CNN, which simultaneously learns to focus on the more discriminative parts of the input scene. Given a global feature, we directly predict the attention mask which is used to re-weight the local scene details in the feature space. This strategy allows the flexibility to re-focus attention on the local details which can be highly valuable for predicting a persons unique identity. The local feature description leads to a highly compact and complementary feature representation, which together with the global representation achieves highly accurate results on three large-scale datasets. Significant boosts are observed when the proposed features are used along-with a re-ranking strategy, demonstrating the strength of proposed features to correctly encode reciprocal relationships between person identities.

References

[1] Barbosa, I.B., Cristani, M., Caputo, B., Rognhaugen, A., Theoharis, T., 2017. Looking beyond appearances: Synthetic training data for deep cnns in re-identification. ArXiv preprint .

[2] Chen, D., Yuan, Z., Chen, B., Zheng, N., 2016. Similarity learning with spatial constraints for person re-identification, in: CVPR.

[3] Chen, W., Chen, X., Zhang, J., Huang, K., 2017a. Beyond triplet loss: a deep quadruplet network for person re-identification, in: CVPR.

[4] Chen, Y., Zhu, X., Gong, S., 2017b. Person re-identification by deep learning multi-scale representations, in: CVPR.

[5] Chen, Y.C., Zhu, X., Zheng, W.S., Lai, J.H., . Person re-identification by camera correlation aware feature augmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[6] Cheng, D., Gong, Y., Li, Z., Zhang, D., Shi, W., Zhang, X., 2018. Cross-scenario transfer metric learning for person re-identification. Pattern Recognition Letters .

[7] Dong, H., Lu, P., Liu, C., Ji, Y., Li, Y., Gong, S., 2018. Person re-identification by kernel null space marginal fisher analysis. Pattern Recognition Letters .

[8] Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M., 2010. Person re-identification by symmetry-driven accumulation of local features, in: CVPR.

[9] Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multiscale, deformable part model, in: CVPR.

[10] Fu, J., Zheng, H., Mei, T., 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition, in: CVPR.

[11] Geng, M., Wang, Y., Xiang, T., Tian, Y., 2016. Deep transfer learning for person re-identification. ArXiv preprint arXiv:1611.05244 .

[12] Guillaumin, M., Verbeek, J., Schmid, C., 2009. Is that you? metric learning approaches for face identification, in: ICCV.

[13] Hansen, L.K., Salamon, P., 1990. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 993–1001.

[14] Hashem, S., 1997. Optimal linear combinations of neural networks. Neural Networks 10, 599–614.

[15] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: CVPR.

[16] Hermans, A., Beyer, L., Leibe, B., 2017. In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 .

[17] Jaderberg, M., Simonyan, K., Zisserman, A., et al., 2015. Spatial transformer networks, in: NIPS.

[18] Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H., 2012. Large scale metric learning from equivalence constraints, in: CVPR.

[19] Krogh, A., Vedelsby, J., 1995. Neural network ensembles, cross validation, and active learning, in: NIPS, pp. 231–238.

[20] Lan, X., Wang, H., Gong, S., Zhu, X., 2017. Deep reinforcement learning attention selection for person re-identification. arXiv preprint arXiv:1707.02785 .

[21] Li, D., Chen, X., Zhang, Z., Huang, K., 2017a. Learning deep context-aware features over body and latent parts for person re-identification, in: CVPR, pp. 384–393.

[22] Li, W., Zhu, X., Gong, S., 2017b. Person re-identification by deep joint learning of multi-loss classification, in: IJCAI.

[23] Li, W., Zhu, X., Gong, S., 2018. Harmonious attention network for person re-identification, in: CVPR, p. 2.

[24] Liang, C., Huang, B., Hu, R., Zhang, C., Jing, X., Xiao, J., 2015. A unsupervised person re-identification method using model based representation and ranking, in: ACM-MM, ACM. pp. 771–774.

[25] Liao, S., Hu, Y., Zhu, X., Li, S.Z., 2015. Person re-identification by local maximal occurrence representation and metric learning, in: CVPR.

[26] Lin, Y., Zheng, L., Zheng, Z., Wu, Y., Yang, Y., 2017. Improving person re-identification by attribute and identity learning. arXiv preprint arXiv:1703.07220 .

[27] Lin, Y., Zheng, Z., Zhang, H., Gao, C., Yang, Y., 2018. Bayesian query expansion for multi-camera person re-identification. Pattern Recognition Letters .

[28] Liu, H., Feng, J., Qi, M., Jiang, J., Yan, S., 2017a. End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing .

[29] Liu, J., Zha, Z.J., Tian, Q., Liu, D., Yao, T., Ling, Q., Mei, T., 2016. Multi-scale triplet cnn for person re-identification, in: ACM-MM, pp. 192–196.

[30] Liu, X., Zhao, H., Tian, M., Sheng, L., Shao, J., Yi, S., Yan, J., Wang, X., 2017b. Hydraplus-net: Attentive deep features for pedestrian analysis. arXiv preprint arXiv:1709.09930 .

[31] Matsukawa, T., Okabe, T., Suzuki, E., Sato, Y., 2016. Hierarchical gaussian descriptor for person re-identification, in: CVPR.

[32] Opitz, D., Maclin, R., 1999. Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11, 169–198.

[33] Opitz, D.W., Shavlik, J.W., 1996. Actively searching for an effective neural network ensemble. Connection Science 8, 337–354.

[34] Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C., 2016. Performance measures and a data set for multi-target, multi-camera tracking, in: ECCV.

[35] Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering, in: CVPR.

[36] Singh, K.K., Lee, Y.J., 2017. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization, in: ICCV.

[37] Sollich, P., Krogh, A., 1996. Learning with ensembles: How overfitting can be useful, in: NIPS, pp. 190–196.

[38] Su, C., Li, J., Zhang, S., Xing, J., Gao, W., Tian, Q., 2017. Pose-driven deep convolutional model for person re-identification, in: ICCV, IEEE. pp. 3980–3989.

[39] Su, C., Zhang, S., Xing, J., Gao, W., Tian, Q., 2016. Deep attributes driven multi-camera person re-identification, in: ECCV.

[40] Sun, Y., Zheng, L., Deng, W., Wang, S., 2017. Svdnet for pedestrian retrieval, in: ICCV.

[41] Ustinova, E., Ganin, Y., Lempitsky, V.S., 2017. Multiregion bilinear convolutional neural networks for person re-identification, in: AVSS.

[42] Varior, R.R., Haloi, M., Wang, G., 2016. Gated siamese convolutional neural network architecture for human re-identification, in: ECCV.

[43] Vedaldi, A., Lenc, K., 2015. Matconvnet – convolutional neural networks for matlab, in: ACM-MM.

[44] Wang, F., Zuo, W., Lin, L., Zhang, D., Zhang, L., 2016. Joint learning of single-image and cross-image representations for person re-identification, in: CVPR.

[45] Wei, L., Zhang, S., Yao, H., Gao, W., Tian, Q., 2017. Glad: global-local-alignment descriptor for pedestrian retrieval, in: Proceedings of the 2017 ACM on Multimedia Conference, ACM. pp. 420–428.

[46] Wu, L., Shen, C., van den Hengel, A., 2017. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. Pattern Recognition 65, 238–250.

[47] Xiao, T., Li, H., Ouyang, W., Wang, X., 2016. Learning deep feature representations with domain guided dropout for person re-identification, in: CVPR.

[48] Xiao, T., Li, S., Wang, B., Lin, L., Wang, X., 2017. Joint detection and identification feature learning for person search, in: CVPR.

[49] Yi, D., Lei, Z., Liao, S., Li, S.Z., 2014. Deep metric learning for person re-identification, in: ICPR.

[50] Zhang, L., Xiang, T., Gong, S., 2016. Learning a discriminative null space for person re-identification, in: CVPR.

[51] Zhao, C., Chen, K., Wei, Z., Chen, Y., Miao, D., Wang, W., 2018. Multilevel triplet deep learning model for person re-identification. Pattern Recognition Letters .

[52] Zhao, R., Ouyang, W., Wang, X., 2013. Unsupervised salience learning for person re-identification, in: CVPR, pp. 3586–3593.

[53] Zheng, L., Huang, Y., Lu, H., Yang, Y., 2017a. Pose invariant embedding for deep person re-identification. CoRR abs/1701.07732.

[54] Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q., 2015. Scalable person re-identification: A benchmark, in: ICCV.

[55] Zheng, L., Yang, Y., Hauptmann, A.G., 2016. Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 .

[56] Zheng, Z., Zheng, L., Yang, Y., 2017b. A discriminatively learned cnn embedding for person re-identification. TOMM .

[57] Zheng, Z., Zheng, L., Yang, Y., 2017c. Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: ICCV.

[58] Zheng, Z., Zheng, L., Yang, Y., 2018. Pedestrian alignment network for large-scale person re-identification. IEEE Transactions on Circuits and Systems for Video Technology .

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