# Dual Attention Matching Network for Context-Aware Feature Sequence based Person Re-Identification

Jianlou Si[1], Honggang Zhang[1], Chun-Guang Li[1], Jason Kuen[2],
Xiangfei Kong[2], Alex C. Kot[2], Gang Wang[3]
[1] Beijing University of Posts and Telecommunications, Beijing, China
[2] Nanyang Technological University, Singapore
[3] Alibaba AI Labs, Hangzhou, China

## Abstract

*Typical person re-identification (ReID) methods usually describe each pedestrian with a single feature vector and match them in a task-specific metric space. However, the methods based on a single feature vector are not sufficient enough to overcome visual ambiguity, which frequently occurs in real scenario. In this paper, we propose a novel end-to-end trainable framework, called Dual ATtention Matching network (DuATM), to learn context-aware feature sequences and perform attentive sequence comparison simultaneously. The core component of our DuATM framework is a dual attention mechanism, in which both intra-sequence and inter-sequence attention strategies are used for feature refinement and feature-pair alignment, respectively. Thus, detailed visual cues contained in the intermediate feature sequences can be automatically exploited and properly compared. We train the proposed DuATM network as a siamese network via a triplet loss assisted with a decorrelation loss and a cross-entropy loss. We conduct extensive experiments on both image and video based ReID benchmark datasets. Experimental results demonstrate the significant advantages of our approach compared to the state-of-the-art methods.*

## 1. Introduction

Person Re-Identification (ReID) aims at associating the same pedestrian across multiple cameras [13, 63], which has attracted rapidly increased attention in the computer vision community due to its importance for many potential applications, such as video surveillance analysis and content-based image/video retrieval. A typical person ReID pipeline usually describes each pedestrian image or video footage with *a single feature vector* firstly and then match them in a task-specific metric space, where the feature vectors of same pedestrian are expected to have smaller distances than that of different pedestrians, *e.g.*, [20, 34, 21, 64, 54]. Recently, benefited from the success of deep learning, feature vector based methods have obtained significant performance improvements [14, 37, 28, 26]. However, when the individuals undergo drastic appearance changes or when they are dressed in similar clothes, it becomes quite difficult to use single feature vector based representation for reliable person ReID. As shown in Fig. 1 (a), different individuals are very similar to each other in appearance, except for some local patterns on skirts. Unfortunately, the single feature vector based methods usually pay more attention on the overall appearance rather than the local discriminative parts and thus fail to yield accurate matching results. Moreover, as shown in Fig. 1 (d), there are also some interference frames in each video sequence, which will heavily contaminate the whole feature vector and thus lead to mismatching.

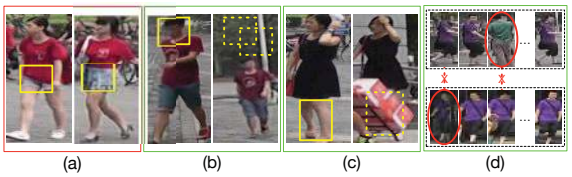



Figure 1. Hard examples in person ReID. (a): negative pair with similar appearance. (b): positive pair with large spatial displacement. (c): positive pair with body-part missing due to occlusion. (d): positive video pair with interference frames (marked by ellipse) and temporal misalignments (indicated by red "×" mark).

An alternative way to address these problems is to describe each person with *a set of feature vectors* and match them based on feature set or feature sequence.[1] For example, in [18, 1, 32, 59, 60], the spatial-patch based local feature sequences or body-part based semantic feature sets are extracted from pedestrian images and matched according to some heuristic correspondence structures; in [44, 43, 69], multiple sub-segment or frame level compar-

[1] In this paper, we refer to a group of feature vectors as feature sequence if they have spatial/temporal adjacent relations; otherwise as feature set.

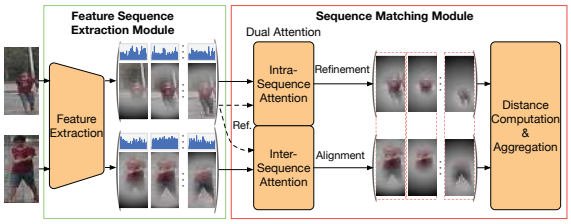



Figure 2. Schematic illustration of DuATM.

isons are computed and aggregated for matching pedestrian based on video. Among these methods, matching the local feature sequences based on a universal correspondence structure might easily fail when encountering heavily sequence misalignments, *e.g.*, caused by the spatial displacements as shown in Fig. 1 (b) or local interferences as shown in Fig. 1 (d). Besides, matching the semantic feature sets based on human body structure might also fail when encountering body-part occlusions as shown in Fig. 1 (c).

To tackle the challenges mentioned above, we propose a novel end-to-end trainable framework, named Dual ATtention Matching network (DuATM), to jointly learn context-aware feature sequences and perform attentive sequences comparison. Our framework consists of two cascaded modules, one for feature sequence extraction and one for feature sequence matching, as illustrated in Fig. 2. The feature sequence extraction module is built on a spatial/bi-recurrent convolutional neural network (CNN) for image/video inputs to extract spatial/temporal-spatial context-aware feature sequences. The sequence matching module is based on a dual attention mechanism—which contains one attention strategy for intra-sequence refinement and one attention strategy for inter-sequence alignment—the former refines each corrupted feature vectors by exploiting the contextual information within sequence and the later aligns feature-pair by selecting semantically consistent counterparts cross paired sequences. After feature sequences refinement and alignment, the holistic sequence distance score is computed by aggregating multiple local distances between the refined and aligned pairwise feature vectors of each paired sequences. We train the proposed DuATM as a siamese network with a triplet loss plus a de-correlation loss and a cross-entropy loss, to reduce the feature sequence redundancy and enhance the feature sequence discrimination.

The main contributions of the paper are as follows.

- We propose a novel end-to-end trainable framework for person ReID, which can jointly learn context-aware feature sequences and perform sequences comparison with dual attention mechanism.

- We use a dual attention mechanism to perform intra-sequence feature refinement and inter-sequence feature-pair alignment simultaneously.

- We train DuATM as a siamese network with a triplet loss, plus a de-correlation loss and a cross-entropy loss, and evaluate the effectiveness of each part.

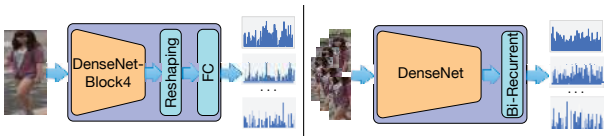
- We conduct extensive experiments on both image and video based benchmark datasets and demonstrate the effectiveness of our proposal.

## 2. Related Works

Person ReID systems usually consist of two major components: a) feature extraction and b) metric learning. Previous works on person ReID focus on either constructing informative features, or finding a discriminative distance metric. According to the used representation forms in matching stage, we roughly divide the existing methods into two groups: *feature vector based methods*, *e.g.*, [10, 4, 19, 37, 41, 17, 12, 45, 35]; and *feature set or feature sequence based methods*, *e.g.*, [69, 60, 59, 57, 36, 18, 1, 38].

In feature vector based methods, an image or video is represented by a feature vector and the metric learning is performed based on feature vectors. For example, in [2, 20, 29, 47, 56, 51, 54, 64, 70, 21], hand-crafted local features are integrated into a feature vector, and distance metric is learned by simultaneously maximizing inter-class margins and minimizing intra-class variations. Meanwhile, many recent works directly learn comparable feature embedding from the raw input data via a neural network. For example, in [33, 26], high-quality local patterns are explored from images or videos firstly and then aggregated into informative feature vectors; in [28, 39, 49], local features of recurrent appearance data are extracted and integrated using temporal-pooling strategy; in [14, 5], to enhance the generalization capability of the learned embeddings, the pairwise similarity criterion is extended to triplet or quadruplet. Although these methods mentioned above are able to learn task-specific compact embeddings, these methods still suffer from the mismatching problem, especially when some vital visual details fail to be captured.

Different from the feature vector based methods, feature set or feature sequence based methods are capable of preserving more detailed visual cues by leveraging complementary feature vectors or spatial information. For example, in [32, 3], local spatial constraints are adopted when computing spatial-patch based feature sets similarity; in [43], dense element-wise correspondences are employed when computing the distance of temporal feature sequences; in [36, 18, 1, 38], spatial correspondence structures are explored via the patch comparison layer inserted in a deep network; in [59, 60], the body structure information is utilized to facilitate the semantic alignment of feature sequences. While these methods mentioned above exploit heuristic correspondence structures to compare feature se-




(a) Image feature extraction (b) Video feature extraction

Figure 3. Feature sequence extraction module.

quences, they might easily fail when heavy misalignments or interferences occur in feature sequences.

Recently, attention mechanism has been proposed in many tasks of matching sequences or learning representations, *e.g.*, [40, 48, 50, 42, 24, 69]. In [40, 42], attention mechanism is used to softly align word embeddings between each text-sequence pair in the task of natural language processing. In [48, 24], glimpse representation is produced for each image via neural attention, so that each input pair can be compared progressively. In [50, 69], fixed-dimension feature vectors are learned from variable length videos for face recognition and person ReID by attentive aggregation, respectively. However, these methods either consider a single intra-sequence attention for feature selection from feature set/sequence, or consider a single inter-sequence attention for feature sets/sequences matching. While inter-sequence attention is able to tackle the sequence misalignment problem, it might fail when interferences or corruptions occur. On the other hand, intra-sequence attention is able to tackle corruptions but it is not able to align sequences.

In this paper, we exploit the attention mechanism into feature sequence based person ReID. Unlike the existing methods, that compare sequences via heuristic correspondence structures, we attempt to compare two sequences via *dual attention processes*, in which an inter-sequence attention process is used to perform sequence alignment and an intra-sequence attention process is used simultaneously to perform sequence refinement.

# 3. Our Proposal: Dual Attention Matching Network (DuATM)

This section will present an end-to-end trainable framework—DuATM, which consists of two modules: one for extracting feature sequences and one for matching feature sequences, as illustrated in Fig. 2.

## 3.1. Feature Sequence Extraction Module

In DuATM, we adopt DenseNet-121 [15] as the backbone of the feature sequence extraction module. Owning to the direct connections between each layer to all the subsequent layers in DenseNet, local details are better propagated to the outputs to enrich the final feature sequences. Specifically, the network architectures for image and video inputs are slightly different.

- Given an image $\mathcal{X} \in \mathbb{R}^{H\times W\times 3}$, as illustrated in Fig. 3 (a), the convolutional feature maps are obtained from the outputs of DenseNet-Block4. Each feature vector at a specific position across all channels contains both the spatial details and semantic contexts due to its large receptive field size. Then, these feature vectors are rearranged by locations to form a feature sequence and each feature vector is further transformed into a compact embedding space via a Fully Connected (FC) layer.

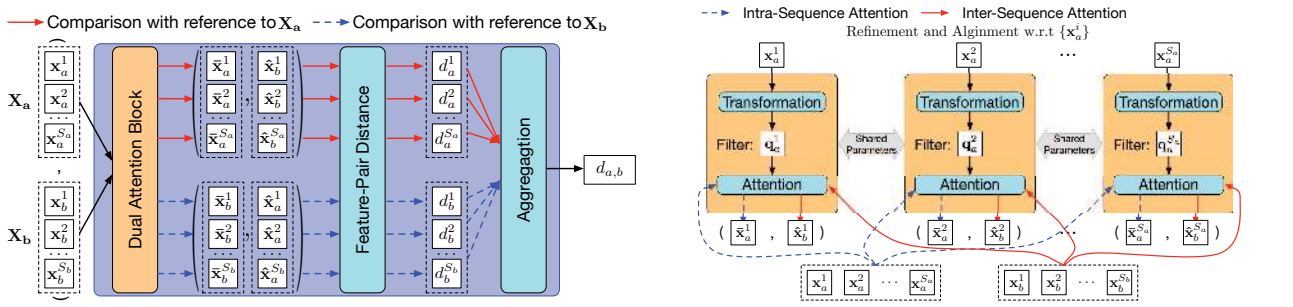
- Given a video footage $\mathcal{X} \in \mathbb{R}^{H\times W\times 3\times T}$, of length $T$, as illustrated in Fig. 3 (b), each frame in video at a time-step is passed to a DenseNet to produce the frame-level feature vector. Then, a bidirectional recurrent layer is introduced to encode both the temporal-spatial appearance details and the complementary motion cues, by allowing information to be passed between time-steps. Finally, all hidden states from different time-steps compose the final feature sequence for the video.

For convenience, we denote the feature sequence extraction as $\mathbf{X} = \mathcal{F}(\mathcal{X}; \Theta_{\mathcal{F}})$, where $\mathbf{X}$ is the extracted feature vectors sequence which encodes spatial or temporal information and $\mathcal{F}(\cdot; \Theta_{\mathcal{F}})$ represents the feature extraction module parameterized with $\Theta_{\mathcal{F}}$. More specifically, we denote $\mathbf{X} = \{\mathbf{x}^i \in \mathbb{R}^D\}_{i=1}^{S}$ as a feature sequence of length $S$. Each feature vector $\mathbf{x}^i$ is normalized to unit $\ell_2$ norm before passing it to the next module.

## 3.2. Sequence Matching Module

Sequence matching module is the most important component of DuATM. Note that there is no supervision information available to force the feature extraction module to learn semantically aligned feature sequences, thus one of the goals of this module is to compare each pair of possibly unaligned feature sequences $(\mathbf{X}_a, \mathbf{X}_b)$, where $\mathbf{X}_a = \{\mathbf{x}_a^i\}_{i=1}^{S_a}$ and $\mathbf{X}_b = \{\mathbf{x}_b^j\}_{j=1}^{S_b}$. However, each sequence may also contain a certain amount of corrupted feature vectors (*e.g.*, caused by the noisy inputs). A naive method is to transform feature sequences into comparable feature vectors via average pooling, in which the misalignment or corruptions are ignored. Instead, we propose to refine and align each feature sequence pair at first, then compute and aggregate the distances of multiple feature pairs.

Since that the intermediate feature sequences obtained from our feature extraction module contain abundant contextual information, we use these contexts to remove the feature corruptions and compare feature sequences. Specifically, we attempt to exploit the contextual information to help feature sequence refinement and feature sequence pair alignment via the attention mechanism. To be more specific, if one of a feature sequence pair is treated as the memory, the refinement of each feature vector within this sequence




(a) Sequence matching module (b) Dual attention block

Figure 4. Illustration for sequence matching module and dual attention block.

can be achieved by an intra-sequence attention process; oppositely, if the other sequence is treated as the memory, the alignment of each feature vector can be achieved by an inter-sequence attention process; and vice versa.

For clarity, we illustrate the sequence matching module of DuATM in Fig. 4 (a), in which two types of attention procedures are integrated into a dual attention block for feature sequence refinement and alignment, as illustrated in Fig. 4 (b). After refinement and alignment, the holistic sequence distance score is computed by aggregating multiple local distances between refined and aligned pairwise feature vectors of each sequence pair.

#### 3.2.1 Dual Attention Block

The dual attention block is composed of one transform layer and one attention layer, in which the transform layer is used to produce the feature-aware filter and the attention layer is used to generate the corresponding attention weights. Without loss of generality, as an example to present the dual attention block in detail, we describe the generation process of $(\bar{\mathbf{x}}_a^i, \hat{\mathbf{x}}_b^i)$, as illustrated in Fig. 4 (b). Specifically, let $\mathbf{x}_a^i$ be the reference feature to be refined and aligned.

- At first, the filter is computed through the transform layer as follows:

$$\mathbf{q}_a^i = ReLU(BN(\mathbf{W}\mathbf{x}_a^i + \mathbf{b})), \tag{1}$$

  where $\mathbf{W}$ and $\mathbf{b}$ are the weight matrix and bias vector of a linear layer, $BN$ and $ReLU$ represent Batch Normalization [16] and rectified linear unit (ReLU) function, respectively.

- Then, the attention significance for intra-sequence refinement and inter-sequence alignment can be computed separately through the attention layer as follows:

$$\bar{e}_a^{i,m} = \langle \mathbf{q}_a^i, \mathbf{x}_a^m \rangle, \quad \hat{e}_b^{i,n} = \langle \mathbf{q}_a^i, \mathbf{x}_b^n \rangle, \tag{2}$$

  where $\langle \cdot, \cdot \rangle$ denotes the inner product.

- Finally, the semantically refined and aligned feature vector pair $(\bar{\mathbf{x}}_a^i, \hat{\mathbf{x}}_b^i)$ is obtained by linearly combining elements within the corresponding sequences, respectively, via normalized attention weights as

$$\bar{\mathbf{x}}_a^i = \sum_{m=1}^{S_a} \sigma(\bar{e}_a^{i,m}) \mathbf{x}_a^m, \quad \hat{\mathbf{x}}_b^i = \sum_{n=1}^{S_b} \sigma(\hat{e}_b^{i,n}) \mathbf{x}_b^n, \tag{3}$$

  where $\sigma(\cdot)$ is a soft-max function for normalization in which $\sigma(t_j) = \frac{\exp(t_j)}{\sum_{j=1}^{S} \exp(t_j)}$ for $\mathbf{t} \in \mathbb{R}^S$.

Following Eq. (1) to (3), the comparable feature pairs with respect to each feature vector can be obtained.

#### 3.2.2 Distance Computation and Aggregation

Owing to the feature sequence refinement and alignment performed in the dual attention block, it is reasonable to directly compute the distance between two refined and simultaneously aligned features, and aggregate the computed distances of feature-pairs into a holistic sequence distance.

In DuATM, the dual attention is bidirectional, *i.e.*, the dual attention process is carried out twice with respect to $\{\mathbf{x}_a^i\}$ and $\{\mathbf{x}_b^j\}$, respectively. Thus, we use the the distances of sequence-pair in both two comparison directions to define the distance of the holistic sequences. Specifically, we use the Euclidean distance to compute the distance between feature pair, *i.e.*,

$$\begin{aligned} d_a^i &= \|\bar{\mathbf{x}}_a^i - \hat{\mathbf{x}}_b^i\|_2, \quad i = 1, \cdots, S_a, \\ d_b^j &= \|\bar{\mathbf{x}}_b^j - \hat{\mathbf{x}}_a^j\|_2, \quad j = 1, \cdots, S_b. \end{aligned} \tag{4}$$

And then, we aggregate these distances via the average-pooling to define the distance of feature sequences $\mathbf{X}_a$ and $\mathbf{X}_b$ as follows:

$$\|\mathbf{X}_a - \mathbf{X}_b\|_{\mathcal{M}} = \frac{1}{2S_a} \sum_{i=1}^{S_a} d_a^i + \frac{1}{2S_b} \sum_{j=1}^{S_b} d_b^j, \tag{5}$$

where $\|\mathbf{X}_a - \mathbf{X}_b\|_{\mathcal{M}}$ is the distance defined by the sequence matching module. For convenience, we denote all parame-

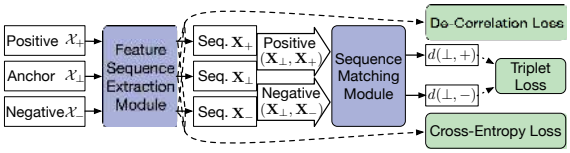



Figure 5. Overall flowchart for training.

ters (*i.e.*, $\mathbf{W}$, $\mathbf{b}$, and the parameters in the BN layer) in the sequence matching module as $\Theta_{\mathcal{M}}$.

### 3.3. Loss Functions for Training DuATM

To train the whole network to perform person ReID and generalize well on unseen data, we use the siamese architecture with *triplet loss* during the training period as shown in Fig. 5. Moreover, to make the learned intermediate feature sequences compact, robust, and more discriminative, we also combined with two auxiliary losses, *i.e. de-correlation loss* and *cross-entropy loss*. Thus, the overall loss function is defined as:

$$\ell = \ell^{(0)}(\mathcal{X}, \Theta_{\mathcal{F}}, \Theta_{\mathcal{M}}) + \lambda_1 \ell^{(1)}(\mathcal{X}, \Theta_{\mathcal{F}}) + \lambda_2 \ell^{(2)}(\mathcal{X}, \Theta_{\mathcal{F}}, \boldsymbol{\theta}),$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are two tradeoff parameters.

**Triplet Loss.** The objective of using triplet loss is to force the network to make the distance between the positive pairs smaller than the negative ones. Given a triplet of person images/videos, the extraction module extracts spatial/temporal-spatial context-aware feature sequences via a three-branch siamese subnet, and the matching module attentively computes the distances between the positive and negative pair via a two-branch siamese subnet.

Let $\mathcal{X} = (\mathcal{X}_\perp, \mathcal{X}_+, \mathcal{X}_-)$ be a triplet input. To force the network to predict the distance of positive pair smaller than the negative pair with a margin $\gamma$, we define the triplet loss $\ell^{(0)}(\mathcal{X}, \Theta_{\mathcal{F}}, \Theta_{\mathcal{M}})$ as:

$$\max\{0, \gamma + \|\mathcal{F}(\mathcal{X}_\perp) - \mathcal{F}(\mathcal{X}_+)\|_{\mathcal{M}} - \|\mathcal{F}(\mathcal{X}_\perp), \mathcal{F}(\mathcal{X}_-)\|_{\mathcal{M}}\}, \tag{6}$$

where $\gamma > 0$ (*e.g.*, $\gamma = 0.2$ in our experiments).

**De-Correlation Loss.** In [9], de-correlating representations has been studied as a regularizer to reduce over-fitting in deep networks. In this paper, we formulate a similar but different de-correlation loss to make our feature sequence more compact. Specifically, we impose a constraint on the intra-sequence correlation matrix as follows:

$$\ell^{(1)}(\mathcal{X}, \Theta_{\mathcal{F}}) = \frac{1}{N^2} ||\mathbf{I} - \mathcal{F}(\mathcal{X})^T \mathcal{F}(\mathcal{X})||_F^2, \tag{7}$$

where $\mathbf{I}$ is an identity matrix and $N$ is the total number of feature vectors in the sequence.

**Cross-Entropy Loss with Data Augmentation.** To learn more informative and robust feature sequences, we also use a cross-entropy loss with data augmentation approach. Specifically, we use the data with the same labels to generate more data for training.

Suppose $\mathbf{X} = \mathcal{F}(\mathcal{X})$ is an intermediate feature sequence, we achieve this goal by first pooling the sequence as $\mathbf{z} = \sum_{i=1}^{S} \omega_i \mathbf{x}_i$, where $\sum_i \omega_i = 1$ and $\omega_i \geq 0$, and then passing the aggregated vector to a FC layer followed by a cross-entropy loss:

$$\ell^{(2)}(\mathcal{X}, \Theta_{\mathcal{F}}, \boldsymbol{\theta}) = -\ln \sigma(\mathbf{w}_c \mathbf{z} + b_c), \tag{8}$$

where $c$ is the same label as the input $\mathcal{X}$, $\{\mathbf{w}_c, b_c\}$ refer to the $c_{\text{th}}$ row of the FC layer's weight matrix and bias vector, respectively, and $\boldsymbol{\theta}$ contains the parameters in the new FC layer. Note that, instead of generating $\mathbf{z}$ by simply average pooling, we propose to introduce a random *convex combination* strategy into the pooling stage by randomly generating $\omega_i \in [0, 1]$ and even reset it to 0 with the probability $p > 0$, but keeping $\sum_i \omega_i = 1$. This can be regarded as a simplified version of the interpolation method [11] to augment training dataset.

## 4. Experiments

To evaluate our proposal, we conduct extensive experiments on three large-scale data sets, including Market-1501 [62], DukeMTMC-reID [65], and MARS [61].

### 4.1. Datasets, Evaluation, and Implementations

**Datasets Description.** Market-1501 is collected from 6 cameras, which contains totally 1,501 identities and 32,668 bounding boxes generated by a DPM-detector. It is split into non-overlapping train/test sets of 12,936/19,732 images as defined in [62], and single-query evaluation mode is adopted in our experiments. DukeMTMC-reID is a subset of DukeMTMC [30] captured with 8 cameras for cross-camera tracking. It includes 1,404 identities, in which one half for training and one half for testing. Specifically, there are 2,228 queries, 17,661 galleries, and 16,522 training images, respectively. MARS is an extension of Market-1501 for video-based ReID. It is composed of 8,298 tracklets for 625 identities for training, and 12,180 tracklets for 636 identities for testing as defined in [61], where the tracklets usually contain 25-50 frames.

**Evaluation Protocol.** For performance evaluation, we employ the standard metrics as in most person ReID literatures, namely the cumulative matching cure (CMC) and the mean Average Precision (mAP). To compute these scores, we re-implement the evaluation code provided by [61] in Python.

**Implementation Details.** We use the DenseNet-121 [15] trained on ImageNet to initialize the DenseNet part in DuATM, and train our network with stochastic gradient descent (SGD) method. To be more specific, we freeze the pre-trained DenseNet parameters and train our model for

| Method & Loss | Market-1501 | | | | DukeMTMC-reID | | | | MARS | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R1 | R5 | R20 | mAP | R1 | R5 | R20 | mAP | R1 | R5 | R20 | mAP |
| AvePool+$\ell^{(0)}$ | 74.20 | 89.67 | 95.58 | 56.88 | 64.05 | 79.44 | 87.52 | 43.79 | 65.45 | 81.92 | 90.10 | 47.26 |
| DuATM+$\ell^{(0)}$ | 79.66 | 91.15 | 96.73 | 63.46 | 68.40 | 81.73 | 89.77 | 48.65 | 66.36 | 83.13 | 90.40 | 48.44 |
| DuATM+$\ell^{(0)}$+$\ell^{(1)}$ | 81.83 | 92.46 | 97.33 | 65.21 | 69.17 | 82.23 | 89.36 | 49.48 | 66.52 | 83.78 | 91.21 | 49.07 |
| DuATM+$\ell^{(0)}$+$\ell^{(2)}$ | 87.50 | 95.37 | 98.01 | 70.02 | 79.40 | 90.04 | 94.25 | 61.55 | 73.74 | 87.73 | 93.84 | 56.36 |
| DuATM+$\ell^{(0)}$+$\ell^{(1)}$+$\ell^{(2)}$ | 88.75 | 95.78 | 98.46 | 70.46 | 81.06 | **91.11** | 95.02 | 62.27 | 74.43 | 89.08 | 94.13 | 58.19 |
| DuATM$^{*}$+$\ell^{(0)}$+$\ell^{(1)}$+$\ell^{(2)}$ | 89.96 | 96.53 | 98.72 | 75.22 | 81.46 | 90.75 | 95.11 | 63.14 | 76.36 | 90.10 | 95.30 | 58.96 |
| DuATM$^{**}$+$\ell^{(0)}$+$\ell^{(1)}$+$\ell^{(2)}$ | **91.42** | **97.09** | **98.96** | **76.62** | **81.82** | 90.17 | **95.38** | **64.58** | **78.74** | **90.86** | **95.76** | **62.26** |

Table 1. Comparison to the baseline model. $^{*}$ We adjust the parameters of loss functions to more appropriate values as obtained in the parameter analysis experiments. $^{**}$ The data augmentation is also adopted during the evaluation stage.

the first 100 epochs, and continue the training of the entire network for other 200 epochs. The learning rate is initialized as 0.01 and changed to 0.001 in the last 50 epochs.

An obstacle in training DuATM with triplet loss is lack of positive pairs compared with negative ones. To alleviate the data imbalance issue, we adopt the hard triplet mining strategy [31, 14] to generate triplet mini-batches. Specifically, each mini-batch contains $P$ persons with $V$ images/tracklets, and all of them are regarded as anchor points to select the corresponding hard positives and negatives. In experiments, we set $(P = 10, V = 4)$ with size $256 \times 128$ for image dataset, and set $(P = 7, V = 3)$ with size $128 \times 64$ for video dataset by default. Besides, we follow the common practices to augment image dataset by using random horizontal flips and random crops [28], and to augment video dataset by randomly selecting video subsequences of 16 consecutive frames.

The dimension $D$ of feature vectors within each sequence is set to 256 for both image and video inputs. Besides, the hyper parameters of loss functions, *i.e.*, $\lambda_1$, $\lambda_2$ and corruption ratio $p$, are set as $\lambda_1 = 0.1, \lambda_2 = 0.5$, and $p = 0.2$ when comparing with the baseline. They are tuned to more proper values in the parameter analysis experiments. During the evaluation, we discard the data augmentation process except when comparing with state-of-the-art methods, and use the sub-sequences of 64 consecutive frames for video ReID.[2] All experiments are implemented with PyTorch on 2 Nvidia Titan-X GPUs.

## 4.2. Evaluations on Performance of DuATM

**DuATM Trained with Different Losses.** To evaluate the contribution of each loss and the dual attention block, we train DuATM and report the results with the following four settings: a) DuATM+$\ell^{(0)}$, b) DuATM+$\ell^{(0)}$+$\ell^{(1)}$, c) DuATM+$\ell^{(0)}$+$\ell^{(2)}$, and d) DuATM+$\ell^{(0)}$+$\ell^{(1)}$+$\ell^{(2)}$. Note that DuATM is built on DenseNet. Thus, as the baseline, we take DenseNet to extract feature sequence, use an average pooling layer to form the holistic feature vector and use Euclidean distance to compare feature vectors. The baseline is trained with the triplet loss $\ell^{(0)}$, denoted as AvePool+$\ell^{(0)}$. Experimental results are presented in Table 1. As observed from Table 1 that, the results of DuATM+$\ell^{(0)}$ consistently outperform that of AvePool+$\ell^{(0)}$ on all three data sets. This confirms the effectiveness of using dual attention block in DuATM: using context-aware feature sequences with dual attentive matching mechanism is more effective than the average-pooling based single feature vector method. The performance is further improved when adding the de-correlation loss $\ell^{(1)}$ and the cross-entropy loss $\ell^{(2)}$. Since that the de-correlation loss $\ell^{(1)}$ does not bring any extra supervision information for discrimination, the performance gain of DuATM+$\ell^{(0)}$+$\ell^{(1)}$ over DuATM+$\ell^{(0)}$ is minor. Interestingly, when the cross-entropy loss is added, the performance is significantly improved. This could be accounted to the supervision information brought by the identity labels. Finally, when combining all three loss functions, the accuracy is further improved.

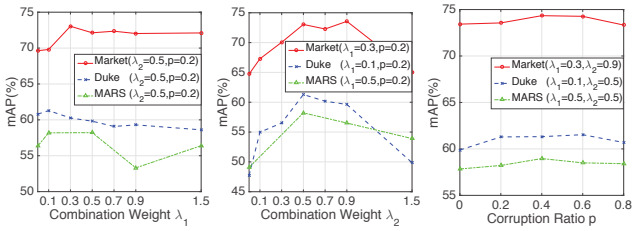
| Method & Loss | R1 | R5 | R20 | mAP |
|---|---|---|---|---|
| AvePool+$\ell^{(0)}$ | 74.20 | 89.67 | 95.58 | 56.88 |
| Intra+$\ell^{(0)}$ | 78.78 | 90.69 | 96.73 | 61.76 |
| Inter+$\ell^{(0)}$ | 72.36 | 87.74 | 95.19 | 53.91 |
| DuATM+$\ell^{(0)}$ | 79.66 | 91.15 | 96.73 | 63.46 |

Table 2. Ablation study of DuATM on Market1501.

**Ablation Study of DuATM.** To verify the effects of intra- and inter-sequence attentions in DuATM, we evaluate each of them separately on Market1501, denoted as Intra+$\ell^{(0)}$ and Inter+$\ell^{(0)}$. Experimental results are listed in Table 2, where DuATM+$\ell^{(0)}$ is nothing but Intra+Inter+$\ell^{(0)}$. We can observe from Table 2 that, jointly using the two attentions, *i.e.*, the dual attention, leads to improvements in the performance than using only one type attention. This confirms the importance of using dual attention mechanism.

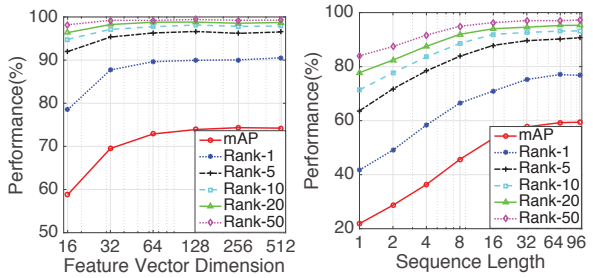
**Evaluation on Parameters in DuATM.** In the loss function of DuATM, there are two parameters $\lambda_1$ and $\lambda_2$. In training the cross-entropy loss, there is also a parameter $p$ to control the corruption ratio in generating auxiliary data. To evaluate the influence of these parameters, we conduct experiments

[2] If the tracklet has less frames, we circularly sample the sequence.




(a) (b) (c)

Figure 6. Evaluation on influence of parameters.




(a) (b)

Figure 7. Evaluation on feature dimension and sequence length.

on three data sets by changing one parameter while fixing the other two. Experimental results are shown in Fig. 6.

From these results, we can draw three conclusions: a) while a moderate value $\lambda_1$ can enforce the sequences more compact, an over-large $\lambda_1$ harms the contextual relationships between feature vectors leading to slightly degenerated performance; b) a moderate value $\lambda_2$ can bring extra supervision information, but an over-large $\lambda_2$ might lead to over-fitting; c) the performance is not sensitive to parameter $p$. Also, we find that DuATM achieves the best performance with the settings of $(\lambda_1, \lambda_2, p)$ as $(0.3, 0.9, 0.4)$, $(0.1, 0.5, 0.6)$, and $(0.5, 0.5, 0.4)$, on Market-1501, DukeMTMC-reID, and MARS, respectively. We list these results in the bottom two rows of Table 1.

In addition, we conduct experiments on Market-1501 and MARS to evaluate the impact of feature dimension $D$ and sequence length $T$ of a video. Experimental results are shown in Fig. 7. For image based ReID, when each person is represented by a sequence with size $D \times T$, even using a lower dimension, the whole sequence can still contain enough discrimination information, *e.g.*, the results at rank-1 still keep at $78.50\%$ or $87.71\%$ for $D = 16$ or $D = 32$, respectively. For video based ReID, since that the feature sequence length is determined by the tracklet size, a longer sequence will contain more visual cues captured at different time points and thus lead to higher accuracy, *e.g.*, mAP is improved from $21.87\%$ with $T = 1$ to $59.42\%$ with $T = 96$.

**Comparison to Other Attention Methods.** To demonstrate the effectiveness of our dual attention mechanism, in Table 3, we compare our method with several existing attentive methods, including CAN [24], HP-Net [25], ST-RNN [69], and QAN [26], in which the salient local patterns are extracted by the attention strategy and aggregated into a single comparable feature vector. Instead, we keep all informative local patterns during feature extraction, and use dual attention mechanism to perform local pattern refinement and pattern-pair alignment during the matching stage. Our model performs a more reasonable comparisons and thus achieves superior performance.

| Dataset | Method | R1 | mAP | Reference |
|---|---|---|---|---|
| Market | CAN | 48.24 | 24.43 | 2017 TIP[24] |
| | HP-Net | 76.90 | - | 2017 ICCV[25] |
| | **DuATM** | **91.42** | **76.62** | This paper |
| MARS | ST-RNN | 70.60 | 50.70 | 2017 CVPR[69] |
| | QAN | 73.74 | 51.70 | 2017 CVPR[26] |
| | **DuATM** | **78.74** | **62.26** | This paper |

Table 3. Comparison to other attention methods.

| Dataset | Method | R1 | mAP | Reference |
|---|---|---|---|---|
| Market | SCSP | 51.90 | 26.35 | 2016 CVPR[3] |
| | SpindleNet | 76.90 | - | 2017 CVPR[59] |
| | DLPAR | 81.00 | 63.40 | 2017 ICCV[60] |
| | DRL-PL | 88.20 | 69.30 | 2017 Arxiv[52] |
| | **DuATM** | **91.42** | **76.62** | This paper |

Table 4. Comparison to other feature sequence/set based methods.

**Comparison to Other Feature Sequence / Feature Set based Methods.** In Table 4, we compare the performance of our method to several existing sequence / set based methods [3, 59, 60, 52]. Since that, DuATM can not only adaptively infer the semantic correspondence structure between local patterns but also automatically remove local corruptions within sequence, our method achieves better performance than body-part based (*e.g.*, SpindleNet [59], DRL-PL [52]) and densely-matching based (*e.g.*, SCSP [3]) methods on dataset Market-1501.

**Comparison to State-of-the-art Methods**[3]**.** In Table 5, Table 6, and Table 7, we compare our approach against the state-of-the-art methods on Market-1501, DukeMTMC-reID, and MARS, respectively. The proposed DuATM achieves superior performance on all of them, that further confirms the effectiveness of our attentively deep context-aware feature sequences based approach. Specifically, in Market-1501 and DukeMTMC-reID, DuATM surpasses all stepwise models and end-to-end networks, and obtains rank-1 accuracy at $91.24\%$ and $81.37\%$ for each dataset. In MARS, DuATM is still better than most approaches above. If our DuATM is trained with a larger image size $256 \times 128$ as in [14], the rank-1 will be $81.16\%$, which surpasses the

[3]Note that different backbones are adopted in different methods, *e.g.*, DenseNet is used in DuATM, ResNet is used in [37, 14], combined multiple networks are used in [19, 67]. Thus, a comprehensive evaluation on the performance with different backbones is a worth future work.

| Method | R1 | R5 | mAP | Reference |
|---|---|---|---|---|
| BOW | 44.42 | 63.90 | 20.76 | 2015 ICCV[62] |
| LDNS | 61.02 | - | 35.68 | 2016 CVPR[55] |
| Re-Rank | 77.11 | - | 63.63 | 2017 CVPR[66] |
| SSM | 82.21 | - | 68.80 | 2017 CVPR[2] |
| S-LSTM | 61.60 | - | 35.30 | 2016 ECCV[39] |
| G-CNN | 65.88 | - | 39.55 | 2016 ECCV[38] |
| CRAFT | 68.70 | - | 42.30 | 2017 TPAMI[8] |
| P2S | 70.72 | - | 44.27 | 2017 CVPR[68] |
| CADL | 73.84 | - | 47.11 | 2017 CVPR[22] |
| USG-GAN | 78.06 | - | 56.23 | 2017 ICCV[65] |
| LDCAF | 80.31 | - | 57.53 | 2017 CVPR[17] |
| SVDNet | 82.30 | 92.30 | 62.10 | 2017 ICCV[37] |
| TriNet | 84.92 | **94.21** | 69.14 | 2017 Arxiv[14] |
| JLML | 85.10 | - | 65.50 | 2017 IJCAI[19] |
| DML | 87.73 | - | 68.83 | 2017 Arxiv[58] |
| REDA | 87.08 | - | 71.31 | 2017 Arxiv[67] |
| DarkRank | **89.80** | - | **74.30** | 2017 Arxiv[6] |
| **DuATM** | **91.42** | **97.09** | **76.62** | This paper |

Table 5. Comparison to state-of-the-art on Market-1501.

| Method | R1 | R5 | mAP | Reference |
|---|---|---|---|---|
| BOW | 25.13 | - | 12.17 | 2015 ICCV[62] |
| LOMO | 30.75 | - | 17.04 | 2015 CVPR[21] |
| USG-GAN | 67.68 | - | 47.13 | 2017 ICCV[65] |
| OIM | 68.10 | - | - | 2017 CVPR[46] |
| APR | 70.69 | - | 51.88 | 2017 Arxiv[23] |
| SVDNet | 76.70 | **86.40** | 56.80 | 2017 ICCV[37] |
| DPFL | 79.20 | - | 60.60 | 2017 ICCVW[7] |
| REDA | **79.31** | - | **62.44** | 2017 Arxiv[67] |
| **DuATM** | **81.82** | **90.17** | **64.58** | This paper |

Table 6. Comparison to state-of-the-art on DukeMTMC-reID.

| Method | R1 | R5 | mAP | Reference |
|---|---|---|---|---|
| SMP | 23.59 | 35.81 | 10.54 | 2017 ICCV[27] |
| BOW | 30.60 | 46.20 | 15.50 | 2015 ICCV[62] |
| DGM | 36.80 | 54.00 | 21.30 | 2017 ICCV[53] |
| Re-Rank | 73.93 | - | 68.45 | 2017 CVPR[66] |
| IDE | 65.10 | 81.10 | 45.60 | 2016 ECCV[61] |
| LDCAF | 71.77 | 86.57 | 56.50 | 2017 CVPR[17] |
| TriNet | **79.80** | **91.36** | **67.70** | 2017 Arxiv[14] |
| **DuATM** | 78.74 | 90.86 | 62.26 | This paper |
| **DuATM*** | **81.16** | **92.47** | **67.73** | This paper |

Table 7. Comparison to state-of-the-art on MARS. DuATM*: trained with a larger image size $256 \times 128$ as suggested in [14].

state-of-the-art result.

**Visualization of Dual Attention Mechanism.** To better understand the dual attention mechanism used in our DuATM, we display some intermediate visualization results in Fig. 8. Since that the learned feature vectors within each sequence are context-aware, with reference to each feature

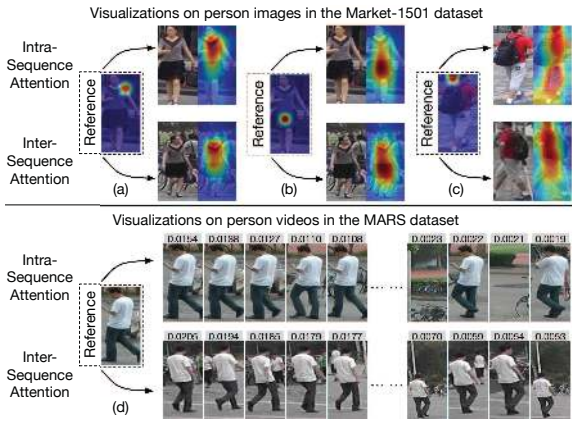



Figure 8. Visualization of the attention weights for intra-sequence and inter-sequence attention, respectively.

vector, the intra-sequence attention can concentrate on its context-related body-parts or gaits from the same image or video to refine itself, and the inter-sequence attention can simultaneously concentrate on the semantically consistent body-parts or gaits from the opposite image or video to generate its aligned counterpart, even when the reference feature is derived from a corrupted region as in Fig. 8 (c). Consequently, in DuATM, the feature sequences are semantically refined and aligned, and thus properly compared.

## 5. Conclusions

We proposed an end-to-end trainable framework, namely Dual ATtention Matching network (DuATM), to learn context-aware feature sequences and to perform dually attentive comparison for person ReID. The core component of DuATM is a dual attention block, which simultaneously performs feature refinement and feature-pair alignment. DuATM is trained via a triplet loss assisted with a de-correlation loss and a cross-entropy loss. Experiments conducted on large-scale image and video data sets have confirmed the significant advantages of our proposal.

## Acknowledgments

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab, Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation and the Infocomm Media Development Authority, Singapore. J. Si, H. Zhang, and C.-G. Li are supported by Beijing Municipal Science and Technology Commission Project under Grant No. Z181100001918005. C.-G. Li is also partially supported by the Open Project Fund from Key Laboratory of Machine Perception (MOE), Peking University. The authors would like to thank the support of the NVIDIA AI Technology Center for their donation/contribution of Titan X GPUs used in our research.

## References


[1] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, pages 3908–3916, 2015. 1, 2

[2] S. Bai, X. Bai, and Q. Tian. Scalable person re-identification on supervised smoothed manifold. In *CVPR*, 2017. 2, 8

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *CVPR*, pages 1268–1277, 2016. 2, 7

[4] J. Chen, Y. Wang, J. Qin, L. Liu, and L. Shao. Fast person re-identification via cross-camera semantic binary transformation. In *CVPR*, 2017. 2

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

[6] Y. Chen, N. Wang, and Z. Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. *arXiv*, 2017. 8

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

[8] Y.-C. Chen, X. Zhu, W.-S. Zheng, and J.-H. Lai. Person re-identification by camera correlation aware feature augmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 8

[9] M. Cogswell, F. Ahmed, R. Girshick, L. Zitnick, and D. Batra. Reducing overfitting in deep networks by decorrelating representations. In *ICLR*, 2016. 5

[10] L. L. A. G. H. Y. G. N. Z. De Cheng, Xiaojun Chang. Discriminative dictionary learning with ranking metric embedded for person re-identification. In *IJCAI*, pages 964–970, 2017. 2

[11] T. DeVries and G. W. Taylor. Dataset augmentation in feature space. *ICLRW*, 2017. 5

[12] M. Geng, Y. Wang, T. Xiang, and Y. Tian. Deep transfer learning for person re-identification. *arXiv*, 2016. 2

[13] S. Gong, M. Cristani, S. Yan, and C. C. Loy. Person re-identification. Springer, 2014. 1

[14] A. Hermans, L. Beyer, and B. Leibe. In defense of the triplet loss for person re-identification. *arXiv*, 2017. 1, 2, 6, 7, 8

[15] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 3, 5

[16] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015. 4

[17] 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

[18] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, pages 152–159, 2014. 1, 2

[19] W. Li, X. Zhu, and S. Gong. Person re-identification by deep joint learning of multi-loss classification. In *IJCAI*, 2017. 2, 7, 8

[20] 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*, pages 3610–3617, 2013. 1, 2

[21] S. Liao, Y. Hu, X. Zhu, and S. Z. Li. Person re-identification by local maximal occurrence representation and metric learning. In *CVPR*, pages 2197–2206, 2015. 1, 2, 8

[22] J. Lin, L. Ren, J. Lu, J. Feng, and J. Zhou. Consistent-aware deep learning for person re-identification in a camera network. In *CVPR*, pages 5771–5780, 2017. 8

[23] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, and Y. Yang. Improving person re-identification by attribute and identity learning. *arXiv*, 2017. 8

[24] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 2017. 3, 7

[25] X. Liu, H. Zhao, M. Tian, L. Sheng, J. Shao, S. Yi, J. Yan, and X. Wang. Hydraplus-net: Attentive deep features for pedestrian analysis. In *ICCV*, 2017. 7

[26] Y. Liu, Y. Junjie, and W. Ouyang. Quality aware network for set to set recognition. In *CVPR*, 2017. 1, 2, 7

[27] Z. Liu, D. Wang, and H. Lu. Stepwise metric promotion for unsupervised video person re-identification. In *ICCV*, 2017. 8

[28] N. McLaughlin, J. Martinez del Rincon, and P. Miller. Recurrent convolutional network for video-based person re-identification. In *CVPR*, pages 1325–1334, 2016. 1, 2, 6

[29] A. Mignon and F. Jurie. Pcca: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, pages 2666–2672, 2012. 2

[30] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCVW*, 2016. 5

[31] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 6

[32] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang. Person re-identification with correspondence structure learning. In *CVPR*, pages 3200–3208, 2015. 1, 2

[33] K. G. Y. Y. S. C. Shuangjie Xu, Yu Cheng and P. Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *ICCV*, 2017. 2

[34] J. Si, H. Zhang, and C.-G. Li. Regularization in metric learning for person re-identification. In *ICIP*, pages 2309–2313, 2015. 1

[35] J. Si, H. Zhang, C.-G. Li, and J. Guo. Spatial pyramid-based statistical features for person re-identification: A comprehensive evaluation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, to appear, 2017. 2

[36] A. Subramaniam, M. Chatterjee, and A. Mittal. Deep neural networks with inexact matching for person re-identification. In *NIPS*, pages 2667–2675, 2016. 2

[37] Y. Sun, L. Zheng, W. Deng, and S. Wang. Svdnet for pedestrian retrieval. In *ICCV*, 2017. 1, 2, 7, 8

[38] R. R. Varior, M. Haloi, and G. Wang. Gated siamese convolutional neural network architecture for human re-identification. In *ECCV*, pages 791–808, 2016. 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*, pages 135–153, 2016. 2, 8

[40] O. Vinyals, S. Bengio, and M. Kudlur. Order matters: Sequence to sequence for sets. In *ICLR*, 2016. 3

[41] 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*, pages 1288–1296, 2016. 2

[42] S. Wang and J. Jiang. A compare-aggregate model for matching text sequences. In *ICLR*, 2017. 3

[43] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by video ranking. In *ECCV*, pages 688–703, 2014. 1, 2

[44] T. Wang, S. Gong, X. Zhu, and S. Wang. Person re-identification by discriminative selection in video ranking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2501–2514, 2016. 1

[45] T. Xiao, H. Li, W. Ouyang, and X. Wang. Learning deep feature representations with domain guided dropout for person re-identification. In *CVPR*, pages 1249–1258, 2016. 2

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

[47] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *ECCV*, pages 1–16, 2014. 2

[48] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 3

[49] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, and X. Yang. Person re-identification via recurrent feature aggregation. In *ECCV*, pages 701–716, 2016. 2

[50] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. In *CVPR*, 2017. 3

[51] Y. Yang, S. Liao, Z. Lei, and S. Z. Li. Large scale similarity learning using similar pairs for person verification. In *AAAI*, pages 3655–3661, 2016. 2

[52] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian. Deep representation learning with part loss for person re-identification. *arXiv*, 2017. 7

[53] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen. Dynamic label graph matching for unsupervised video re-identification. In *ICCV*, 2017. 8

[54] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *CVPR*, pages 1345–1353, 2016. 1, 2

[55] L. Zhang, T. Xiang, and S. Gong. Learning a discriminative null space for person re-identification. In *CVPR*, pages 1239–1248, 2016. 8

[56] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan. Sample-specific svm learning for person re-identification. In *CVPR*, pages 1278–1287, 2016. 2

[57] Y. Zhang, X. Li, L. Zhao, and Z. Zhang. Semantics-aware deep correspondence structure learning for robust person re-identification. In *IJCAI*, pages 3545–3551, 2016. 2

[58] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. Deep mutual learning. *arXiv*, 2017. 8

[59] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. CVPR, 2017. 1, 2, 7

[60] L. Zhao, X. Li, J. Wang, and Y. Zhuang. Deeply-learned part-aligned representations for person re-identification. In *ICCV*, 2017. 1, 2, 7

[61] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *ECCV*, 2016. 5, 8

[62] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015. 5, 8

[63] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv*, 2016. 1

[64] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, 2013. 1, 2

[65] Z. Zheng, L. Zheng, and Y. Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017. 5, 8

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

[67] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang. Random erasing data augmentation. *arXiv*, 2017. 7, 8

[68] 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

[69] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan. See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification. In *CVPR*, 2017. 1, 2, 3, 7

[70] X. Zhu, X.-Y. Jing, F. Wu, and H. Feng. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. In *IJCAI*, pages 3552–3559, 2016. 2