ALL FOR ONE: FRAME-WISE RANK LOSS FOR IMPROVING VIDEO-BASED PERSON RE-IDENTIFICATION

Navaneet K L, Vasudha Todi, R. Venkatesh Babu and Anirban Chakraborty

Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India

ABSTRACT

Person re-identification involves retrieving correct matches for a target image (query) from a set of gallery images, while video based re-identification extends this to the case of query and gallery videos. Typical video-based re-id methods ignore the temporal evolution of the intermediate representations of the video sequences. We propose a novel loss function, termed *rank loss*, to explicitly ensure that the learnt representations achieve enhanced performance and robustness as the sequence progresses and that better intermediate representations result in an improved final representation. Experiments indicate that the addition of rank loss indeed helps in improving the re-id performance while achieving performance comparable to state-of-the-art approaches.

Index Terms— Person re-identification, Rank loss, Attention, Recurrent networks, Video re-identification

1. INTRODUCTION

Person re-identification (re-id) is the task of finding a suitable match for a probe image/video of a person from a set of gallery images/videos. It has important applications in the field of video surveillance, multi-camera person recognition and tracking. The task of re-id proves to be interesting and challenging due to large variations in appearance, pose and illumination across camera fields-of-view. Existing works in person re-id can broadly be classified into three groups - feature learning based [1, 2, 3, 4], metric learning based [5, 6, 7, 8, 9] and deep learning based where feature and metric are jointly learnt [10, 11, 12, 13]. In video based person re-id, instead of comparing single images per target, a sequence of images are considered for each person in the probe and gallery sets. Typically, the individual features of the images in the sequence are obtained and are fused to obtain a single representation for the entire video [12, 14, 15]. The problem then reduces to that of traditional metric learning in person re-id. A number of functions have been tried towards aggregation of features. Zheng et al. [12] use mean and max pooling of features as the fusion function. In [14], McLaughlin et al. use a recurrent neural network (RNN) to learn a fusion function, while Yan et al. [15] use LSTM, a variant of RNN, to learn a more selective composition.

In most classical deep metric learning approaches to video-based person re-id, the attempt has been to obtain an optimal representation after aggregating all the observations in each sequence. Hence, only the final representation is utilized during training the network while the temporally intermediate representations are ignored. A sequence of observations from a target is expected to contain more information on the target's appearance than that contained in any of its subsets. Thus an ideal fusion function, while combining these observations in a sequence, would yield better fused representations as more and more observations are added to this sequence. Our objective in this work, therefore, is to model the fusion function such that it not only is capable of generating an optimal final representation after fusion of all features, but also learns to yield monotonic improvements at each intermediate representation. In the training stage, this desired property must be explicitly enforced through a novel and advanced loss function. However, designing objective functions to obtain such intermediate fused representations is non-trivial, since there is no correspondence between the respective sub-sequences.

We propose a novel loss formulation for the task of video-based person re-id, called *rank loss*, to ensure that the fused representation improves in quality as more information is added, while preventing any degradation due to adverse frames. The network is penalized if, upon the inclusion of a new frame, the fused representation is worse off than any of the previous fused representations. To ensure that the network is not wrongly penalized when a frame with relatively low new information or spurious content is input, a residue based temporal attention network is employed. This helps in enforcing the rank loss on just the relatively clean frames. We demonstrate that the proposed loss, along with the attention network, helps in obtaining a more robust representation for a video and demonstrate that the proposed training strategy helps in improving the retrieval performance.

2. APPROACH

Let $\{s^i_{1,j}, s^i_{2,j}, \ldots, s^i_{T,j}\} \in S^i_j$ be the j^{th} input image sequence of person identity i. A convolutional neural network is employed to extract features from the individual frames of the sequence. Let $\{x^i_{1,j}, x^i_{2,j}, \ldots, x^i_{T,j}\}$ be the corresponding

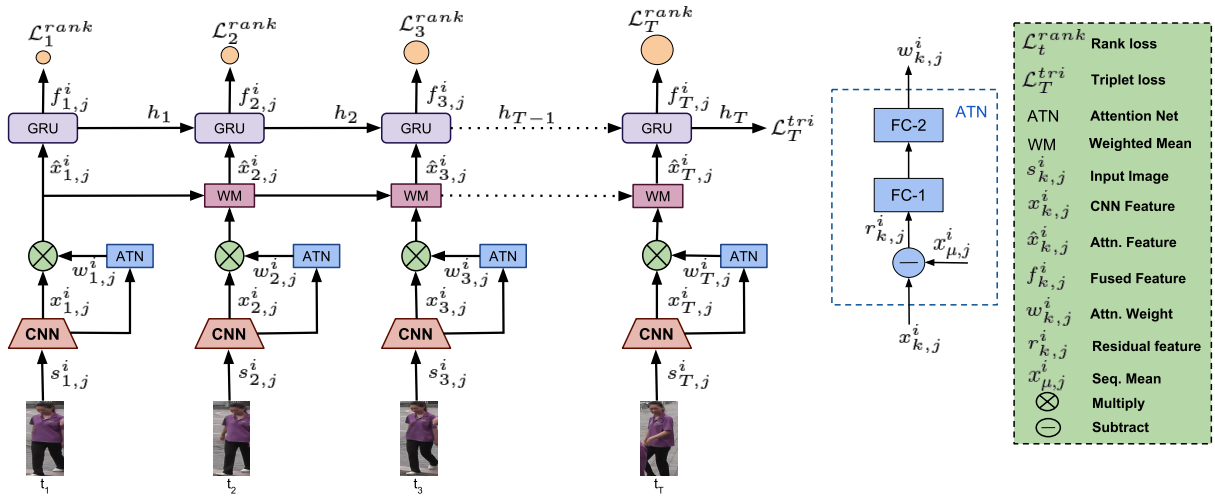

Fig. 1: Overview of proposed approach. Individual frame features are extracted using a CNN. The attention network calculates the residual feature and provides importance scores for each frame. The CNN features are multiplied by attention weights and are then fused using a GRU. Rank loss is applied on the fused representations at every time step, with the circle size indicating weight of the loss at each step. Triplet loss is enforced only on the final representation.

outputs of the convolutional network.

2.1. Temporal Feature Fusion

In order to obtain a combined representation for all the individual frames in the video, temporal pooling is necessary. Unlike simple functions like mean and max pooling which ignore the sequence ordering, we desire a fusion function which can model the temporal evolution of the sequence, while being able to handle input sequences of arbitrary length. We employ gated recurrent unit (GRU), a popular recurrent network variant, to achieve this (Fig. 1). The GRU is used to transform the sequence of CNN features $\{x_1, x_2, \ldots, x_T\}$ (the identity and video indices are dropped for notational ease) from the T individual frames to a corresponding sequence of fused representations $\{f_1, f_2, \ldots, f_T\}$. This is achieved by obtaining a lower dimensional embedding for the input feature, and combining the embedding with the existing state representation of the GRU. Specifically, the following set of transformations are applied at each index t of the sequence:

$$r_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1} + b_r) \tag{1a}$$
$$z_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1} + b_z) \tag{1b}$$
$$s_t = \tanh(W_{hx}x_t + W_{hh}(h_{t-1} \odot r_t) + b_h) \tag{1c}$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot s_t \tag{1d}$$

Here, $\odot$ represents element-wise multiplication, σ represents the sigmoid function. We consider h_t as the pooled feature representation f_t for the input feature sequence $\{x_1, x_2, \ldots, x_t\}$ seen until that point. Internal gating of the new input and its transformed embedding, represented by the intermediate transformations r_t, s_t and z_t, allow the network to selectively update its internal state representation. The network can learn to explicitly consider only the relevant information and ignore the rest. Note that the subscripted W's and b's are shared across all the sequence indices and form the trainable parameters of the GRU.

We employ triplet loss to train the network. Consider an anchor sequence S_j^i of identity i. Any other sequence with the same identity i is considered a positive sequence while the rest of the sequences, with any other identity, are termed negative sequences. The triplet loss tries to ensure that the distance between the anchor and the positive sample is less than the distance between the anchor and the negative sample by a pre-fixed margin. The loss is applied at the final index T of the GRU, with anchor, positive and negative features being $f_{T,j}^i$, $f_{T,k}^i$ and $f_{T,l}^n$ respectively:

$$\mathcal{L}_T^{tri} = \max_{\substack{m \neq i \\ k \neq j}}(0, d(f_{T,j}^i, f_{T,k}^i) - d(f_{T,j}^i, f_{T,l}^n) + m) \tag{2}$$

Here, $d(.)$ is the Euclidean distance metric and m is the pre-fixed margin. In the training stage, the positive and negative samples are chosen using hard-mining within a mini-batch [16].

2.2. Rank Loss

Given the input feature sequence $\{x_{1,j}^i, x_{2,j}^i, \ldots, x_{T,j}^i\}$, the GRU is used to obtain the fused sequence of representations $\{f_{1,j}^i, f_{2,j}^i, \ldots, f_{T,j}^i\}$. While the triplet loss aims to improve the final fused feature at index T, we aim to learn a fused representation $f_{t,j}^i$ at time index t such that it is better than all the fused representations of indices preceding it, i.e, $f_{t,j}^i$ must be a more holistic representation of the sub-sequence till t than any of the representations in $\{f_{1,j}^i, f_{2,j}^i, \ldots, f_{t-1,j}^i\}$.

To achieve such a desirable monotonic improvement in the learnt intermediate representations, we propose the fol-

lowing novel loss function on the outputs of the GRU:

$$\mathcal{L}_t^p = \sum_{i,j} \max_{\substack{\tau \in \{1,2,\ldots,T\} \\ l \in \{1,2,\ldots,K^i\}}} (0, d(f_{t,j}^i, f_{\tau,l}^i) - d_{t,p}^i) \tag{3a}$$

$$\mathcal{L}_t^n = \sum_{i,j} \max_{\substack{\tau \in \{1,2,\ldots,T\} \\ l \in \{1,2,\ldots,K^k\} \\ k \neq i}} (0, d_{t,n}^i - d(f_{t,j}^i, f_{\tau,l}^k)) \tag{3b}$$

$$\mathcal{L}_t^{rank} = \mathcal{L}_t^p + \mathcal{L}_t^n \tag{3c}$$

Here, K^i is the total number of sequences belonging to identity i, T is the maximum sequence length and the indices k and l are chosen using hard-mining. $d_{t,p}^i$ and $d_{t,n}^i$ are defined as follows:

$$d_{t,p}^i = \min_{l \in \{1,2,\ldots,K^i\}, l \neq j} d(f_{\tau,j}^i, f_{r,l}^i) \tag{4a}$$

$$d_{t,n}^i = \max_{l \in \{1,2,\ldots,K^k\}, k \neq i} d(f_{\tau,i}, f_{r,l}^k) \tag{4b}$$

where $\tau \in \{1, 2, \ldots, t-1\}, r \in \{1, 2, \ldots, T\}$. $\mathcal{L}_t^p$ and $\mathcal{L}_t^n$ correspond to the loss on positive and negative samples respectively. As in the case of triplet loss, the positive and negative samples are obtained using hard-mining. Thus, at a give time index t, $d_{t,p}^i$ is the minimum distance between any of the previous fused features till t and the corresponding hard-mined positive sample. $\mathcal{L}_t^p$ tries to ensure that the maximum distance between the current feature and the positive sequence features is less than $d_{t,p}^i$. Similarly, $d_{t,n}^i$ is the maximum distance between any of the previous fused features and the corresponding negative sample. $\mathcal{L}_t^n$ is used to enforce the current feature to be farther from the negative sample than $d_{t,n}^i$. The sum of the two loss functions, termed as rank loss, helps in obtaining an improvement in the fused representation with additional information.

The total loss for training is a combination of triplet and rank loss averaged over all time indices. Within a sequence, more weightage is given to rank loss at later time indices. Specifically, we use a linear weighting scheme, with the weights ranging from zero to one, i.e., $\gamma_t = t/T$. The final loss equation is as follows:

$$\mathcal{L} = \mathcal{L}_T^{tri} + \lambda \sum_{t=1}^{T} \gamma_t \mathcal{L}_t^{rank} \tag{5}$$

2.3. Temporal Attention

While the rank loss expects the fused representation to improve as the sequence progresses, the new inputs need not necessarily provide helpful information. The additional frames might be highly correlated with the currently seen frames, the region of interest might be occluded or the frame could include distractions like the presence of additional person/s. In such situations, it would be wrong to expect the network to improve upon the performance of the existing representation. Thus, to alleviate this issue, an attention network is used to determine the importance of the input frame. Instead of the feature corresponding to just the current frame, a weighted average of the sequence till time t is used as the input to the GRU at time t. The weights for the feature of each frame are learnt through the temporal attention network. Let $\{w_{1,j}^i, w_{2,j}^i, \ldots, w_{t,j}^i\}$ be the weights of the input sequence till time t. The attention network is modelled as a multi-layer perceptron (MLP) with two hidden layers, with ReLU activation as the non-linear function.

$$r_{t,j}^i = x_{t,j}^i - \frac{1}{T} \sum_{\tau=1}^{T} x_{\tau,j}^i \tag{6a}$$

$$w_{t,j}^i = \mathrm{W}_{att}(\mathrm{relu}(\mathrm{U}_{att} r_{t,j}^i + \mathrm{b}_{att}^1)) + \mathrm{b}_{att}^2 \tag{6b}$$

where $W_{att}, U_{att}, b_{att}^1, b_{att}^2$ are the weights and biases of the network. Note that the weights are fixed for all time indices of the sequence (Fig. 1). The residual feature is used for attention calculation. It is obtained as a difference between the current input feature and the average feature of all time steps (Eq. 6a). The residual denotes the extent of additional information provided by the frame and the attention network determines whether the additional information is beneficial. We find this modification to be crucial to effectively learn the attention weights. The input to the GRU at time step t is now given by: $\hat{x}_{t,j}^i = \frac{1}{\sum_{k=1}^{t} w_{k,j}^i} \sum_{\tau=1}^{t} w_{\tau,j}^i * x_{\tau,j}^i$

3. EXPERIMENTS

We perform experimental analysis on two popular video re-id datasets, namely, PRID-2011 [17] and MARS [18]. We consider a Resnet-50 architecture based feature extraction network and use GRU for temporal feature fusion. Triplet loss is used for training the network, without the additional attention module or rank loss. We refer to this network, trained only on triplet loss, as our baseline network. To validate our contributions, we train the same model with an additional rank loss and an attention module which is referred to as baseline+attn+rank loss. We consider rank-1 accuracy and mean average precision (mAP) for evaluation.

3.1. Implementational Details

We use a fully connected layer after the Resnet-50 network to reduce the embedding dimension to 512. Dropout with rate 0.5 is used in this layer. The GRU hidden state size is set to 128 and 512 respectively for PRID-2011 and MARS datasets. The networks are trained end-to-end for 15000 iterations with Adam optimizer, with an initial learning rate of 0.0001, and β_1 and β_2 parameters set to 0.9 and 0.999 respectively. Learning rate scheduling is performed as in [16].

Table 1: Comparison of rank-1 accuracies and mAP for the baseline and proposed approaches on MARS dataset.

Approach	Rank-1	mAP
Baseline	75.51	63.51
Baseline + Attn	76.06	63.20
Baseline + Attn + Rank Loss	**77.27**	**64.76**

3.2. Results on MARS Dataset

MARS [18] dataset has tracklets from six cameras with 1261 identities appearing in a minimum of two cameras. There are 625 identities and 8298 tracklets in the train set while the test set consists of 636 identities and 12180 tracklets. Table 1 provides the quantitative comparison of the baseline with the proposed variants. We observe that inclusion of attention individually helps in slightly improving the rank-1 accuracy, however, the best performance is obtained when the attention network is combined with rank loss based training, emphasizing the need for such an approach. Table 2 provides comparison with the existing approaches. We outperform most of the approaches in Rank-1 and all the approaches in mAP. Note that several existing approaches, like [18], employ metric learning atop their trained model to significantly boost the performance. The approaches also utilize complex modules apart from the basic CNN feature extractor to improve the performance. However, the aim of our work is not to compete against such complex approaches, but to show that the rank loss trained model outperforms the baseline. The proposed training strategy can easily be integrated with the existing approaches to potentially enhance their retrieval performance.

3.3. Results on PRID-2011

PRID-2011 [17] dataset consists of 400 image sequences for 200 identities from two non-overlapping cameras. The sequence lengths range from 5 to 675 frames. Following the protocol in [19], 178 sequences, each with more than 21 frames are considered in our experiments. Since training a deep neural network on such a small dataset might result in overfitting, we use the network pre-trained on MARS and fine-tune it on PRID-2011 for 5000 iterations for both baseline and our approach. Evaluation is done on 10 different train and test splits and the averaged metrics are reported (Table 2). We observe that the addition of rank loss clearly improves the retrieval performance. We also outperform or perform comparably to the existing approaches.

3.4. Role of Rank loss

Tables 1 and 2 provide quantitative evidence for the efficacy of rank loss. Fig. 2 shows a plot of rank-1 accuracy as a function of time-step when the network is trained with and without the rank loss. We observe an increasing trend for both the approaches with increase in the number of input images. The rank loss trained model consistently outperforms the baseline, indicating improved intermediate representations, with the difference between the approaches increasing at the higher time-indices where more images are fused.

Table 2: Comparison of rank-1 accuracies and mAP on MARS and PRID-2011 datasets.

Approach	PRID-2011	MARS	
	Rank-1	Rank-1	mAP
STA [20]	64	-	-
AFDA [21]	43	-	-
RFA [15]	58.2	-	-
CNN+XQDA [18]	77.3	68.3	49.3
CNN-RNN [14]	70	-	-
ASTPN [22]	77	-	-
IDE+XQDA [23]	-	65.3	47.6
MSCAN+Euclidean[24]	-	78.28	61.62
Baseline + Attn	71.46	76.11	63.20
Baseline + Attn + R-Loss	75.17	77.27	64.76

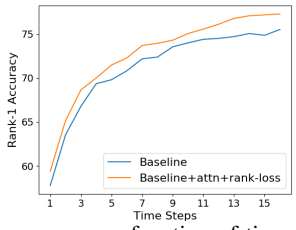

Fig. 2: Rank-1 accuracy as a function of time-step. Training with rank loss achieves higher improvement over the baseline as the sequence progresses.

3.5. Role of Attention

Fig. 3 displays the learnt attention weights for some sample sequences. We observe that the frames with occlusions and those where some body parts are not visible are given a lower score while those with clear views and discriminative features obtain higher scores. For e.g., in row one, notice that the weights decrease in the presence of an occlusion, and increase again when the occlusion disappears.

Fig. 3: Example video sequence with the predicted attention weights for each frame.

4. CONCLUSION

We considered the task of video based person re-identification. A CNN and GRU were used for feature extraction and fusion respectively. To improve the fusion as one observes more frames of a sequence, a novel loss function called rank loss was proposed. A residual input based attention network was employed to determine the relative importance of an input frame. Through extensive experiments on two video re-id datasets, we validated the efficacy of the proposed attention module and rank loss based training strategy. The proposed approach outperforms the baseline on both the datasets.

5. REFERENCES

[1] Douglas Gray and Hai Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," *ECCV*, pp. 262–275, 2008. 1

[2] Igor Kviatkovsky, Amit Adam, and Ehud Rivlin, "Color invariants for person reidentification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1622–1634, 2013. 1

[3] Yang Hu, Shengcai Liao, Zhen Lei, Dong Yi, and Stan Li, "Exploring structural information and fusing multiple features for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 794–799. 1

[4] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin, "Person re-identification: What features are important?," in *European Conference on Computer Vision*. Springer, 2012, pp. 391–401. 1

[5] Zhen Li, Shiyu Chang, Feng Liang, Thomas S Huang, Liangliang Cao, and John R Smith, "Learning locally-adaptive decision functions for person verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3610–3617. 1

[6] Bryan James Prosser, Wei-Shi Zheng, Shaogang Gong, Tao Xiang, and Q Mary, "Person re-identification by support vector ranking.," in *British Machine Vision Conference*, 2010, vol. 2, p. 6. 1

[7] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2288–2295. 1

[8] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3318–3325. 1

[9] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2197–2206. 1

[10] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3908–3916. 1

[11] Rahul Rama Varior, Mrinal Haloi, and Gang Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European Conference on Computer Vision*. Springer, 2016, pp. 791–808. 1

[12] Liang Zheng, Yi Yang, and Alexander G Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016. 1

[13] Zhedong Zheng, Liang Zheng, and Yi Yang, "A discriminatively learned cnn embedding for person re-identification," *arXiv preprint arXiv:1611.05666*, 2016. 1

[14] Niall McLaughlin, Jesus Martinez del Rincon, and Paul Miller, "Recurrent convolutional network for video-based person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1325–1334. 1, 4

[15] Yichao Yan, Bingbing Ni, Zhichao Song, Chao Ma, Yan Yan, and Xiaokang Yang, "Person re-identification via recurrent feature aggregation," in *European Conference on Computer Vision*. Springer, 2016, pp. 701–716. 1, 4

[16] Alexander Hermans*, Lucas Beyer*, and Bastian Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv preprint arXiv:1703.07737*, 2017. 2, 3

[17] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian conference on Image analysis*. Springer, 2011, pp. 91–102. 3, 4

[18] Liang Zheng, Zhi Bie, Yifan Sun, Jingdong Wang, Chi Su, Shengjin Wang, and Qi Tian, "Mars: A video benchmark for large-scale person re-identification," in *ECCV*, 2016. 3, 4

[19] Taiqing Wang, Shaogang Gong, Xiatian Zhu, and Shengjin Wang, "Person re-identification by video ranking," in *ECCV*, 2014. 4

[20] Kan Liu, Bingpeng Ma, Wei Zhang, and Rui Huang, "A spatio-temporal appearance representation for viceo-based pedestrian re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3810–3818. 4

[21] Yang Li, Ziyan Wu, Srikrishna Karanam, and Richard J Radke, "Multi-shot human re-identification using adaptive fisher discriminant analysis.," in *BMVC*, 2015, vol. 1, p. 2. 4

[22] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *CVPR*, 2017. 4

[23] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, Qi Tian, et al., "Person re-identification in the wild.," in *CVPR*, 2017, vol. 1, p. 2. 4

[24] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *CVPR*, 2017. 4