PERSON RE-IDENTIFICATION USING VISUAL ATTENTION

Alireza Rahimpour, Liu Liu, Ali Taalimi, Yang Song, Hairong Qi

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville, TN, USA 37996
{arahimpo, lliu25, ataalimi, ysong18, hqi}@utk.edu

ABSTRACT

Despite recent attempts for solving the person re-identification problem, it remains a challenging task since a person's appearance can vary significantly when large variations in view angle, human pose and illumination are involved. The concept of attention is one of the most interesting recent architectural innovations in neural networks. Inspired by that, in this paper we propose a novel approach based on using a gradient-based attention mechanism in deep convolution neural network for solving the person re-identification problem. Our model learns to focus selectively on parts of the input image for which the networks' output is most sensitive to. Extensive comparative evaluations demonstrate that the proposed method outperforms state-of-the-art approaches, including both traditional and deep neural network-based methods on the challenging CUHK01 and CUHK03 datasets.

Index Terms— Person Re-identification, Gradient-based Attention Network (GAN), Triplet loss, Deep CNN.

1. INTRODUCTION

Recently, person re-identification has gained increasing research interest in the computer vision community due to its importance in multi-camera surveillance systems. Person re-identification is the task of matching people across non-overlapping camera views at different times. A typical re-identification system takes as input two images of person's full body, and outputs either a similarity score between the two images or the decision of whether the two images belong to the same identity or not. Person re-identification is a challenging task. In fact, different individuals can share similar appearances and also appearance of the same person can be drastically different in two different views due to several factors such as background clutter, illumination variation and pose changes.

It has been proven that humans do not focus their attention on an entire scene at once when they want to identify another person [1]. Instead, they *pay attention* to different parts of the scene (e.g., the person's face) to extract the most discriminative information. Inspired by this observation, we study the impact of visual attention in solving person re-identification problem. The visual attention mechanism can significantly reduce the complexity of the person re-identification task, where the network learns to focus on the most informative regions of the scene and ignores the irrelevant parts such as background clutter. Exploiting the attention mechanism in person re-identification task is also beneficial at scaling up the system to large high quality input images. With the recent surge of interest in deep neural networks, attention based models have been shown to achieve promising results on several challenging tasks, including caption generation [1] and machine translation [2] as well as object recognition [3]. However, attention models proposed so far, require defining an explicit predictive model, whose training can pose challenges due to the non-differentiable cost. Furthermore, these models employ Recurrent Neural Network (RNN) for the attention network and are computationally expensive or need some specific policy algorithms such as REINFORCE [3, 4] for training. In this paper we introduce a novel model architecture for person re-identification task which improves the matching accuracy by taking advantage of visual attention. The contributions of this paper are the following:

- We propose a CNN-based task-driven attention model which is specifically tailored for the person re-identification task in a triplet architecture.
- The proposed gradient-based attention model is easy to train and the whole network can be trained in an end-to-end manner with backpropagation and it does not require a policy network (such as reinforcement learning [3]) for training.
- The network is computationally efficient since it first finds the most discriminative regions in the input image and then performs the deep CNN feature extraction only on these selected regions.
- Finally, we quantitatively validate the performance of our proposed model by comparing it to the state-of-the-art performance on two challenging benchmark datasets: CUHK01 [5] and CUHK03 [6].

2. RELATED WORK

Generally, existing approaches for person re-identification are mainly focused on two aspects: learning a distance

metric [7–13] and developing a new feature representation [14–25]. In distance metric learning methods, the goal is to learn a metric that emphasizes inter-personal distance and de-emphasizes intra-person distance. The learnt metric is used to make the final decision as to whether a person has been re-identified or not. Several metric learning algorithms such as Keep It Simple and Straightforward Metric Learning (KISSME) [26], Cross-view Quadratic Discriminant Analysis (XQDA) [7], Metric Learning with Accelerated Proximal Gradient (MLAPG) [12] and Local Fisher Discriminant Analysis (LFDA) [10] were proposed for person re-identification, achieving remarkable performance in several benchmark datasets. In the second group of methods based on developing new feature representation for person re-identification, novel feature representations were proposed to address the challenges such as variations in illumination, pose and view-point [14]. The Scale Invariant Local Ternary Patterns (SILTP) [17], Local Binary Patterns (LBP) [19], Color Histograms [20] or Color Names [23] (and combination of them), are the basis of the majority of these feature representations developed for human re-identification.

In the recent years, several approaches based on Convolutional Neural Network (CNN) architecture for human re-identification have been proposed and achieved great results [6, 27, 28]. In most of the CNN-based approaches for re-identification, the goal is to jointly learn the best feature representation and a distance metric (mostly in a Siamese fashion [29]). With the recent development of RNN networks, the attention-based models have demonstrated outstanding performance on several challenging tasks including action recognition [30]. Except for one recent work [31], the attention mechanism has not yet been studied in the person re-identification literatures. In [31], the RNN-based attention mechanism is similar to the attention model introduced in [30] for action recognition. Different from [31] and inspired by the dynamic model introduced in [32], in our model, the selection of the salient regions is made using a novel gradient-based attention mechanism, that efficiently identifies the input regions for which the network's output is most sensitive to. Moreover, our CNN-based model does not use the RNN architecture as in [31], thus is computationally more efficient and easier to train. Furthermore, in [31] the attention model requires a set of multiple glimpses to estimate the attention which is not required in our proposed architecture.

3. MODEL ARCHITECTURE

In this section we introduce our gradient-based attention model within a triplet comparative platform specifically designed for the task of person re-identification. We first describe the overall structure of our person re-identification design, then we elaborate on the network architecture of the proposed attention mechanism.

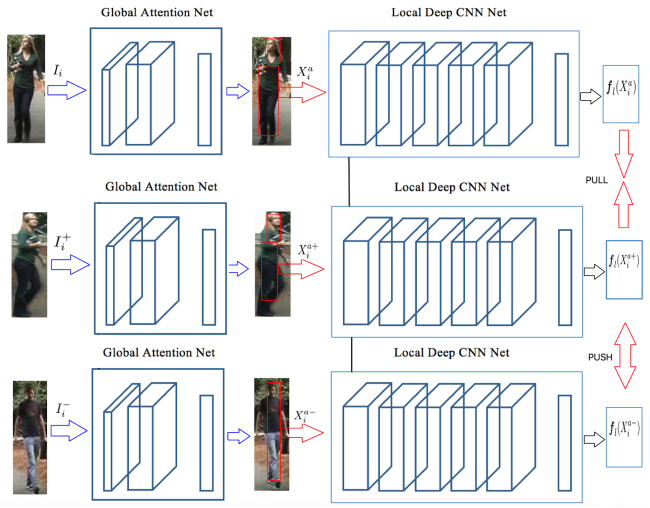

Fig. 1: The architecture of the proposed Gradient-based Attention Network (GAN) in training phase. First the global attention network is trained on input images. Then the Local deep CNN network is trained only over the attended regions of the input image I_i determined by the attention network (i.e., X_i^a, shown with red boxes). The triplet loss forces the distance between feature representation of image pairs of the same person to be minimum and the distance of negative feature pairs to be maximum (See Eq.1 for more details). Best viewed in color.

3.1. Triplet Loss

During training, the three CNNs (with shared parameter set) take the triplets of images $< I_i^+, I_i^-, I_i >$ as input, where I_i^+ and I_i are images from the same person and I_i^- is the image from a different person. As illustrated in Figure 1, each image initially goes through the global attention network and salient regions of the image are selected (i.e., X^a). Then only these selected regions of the image pass through the local deep CNN. The local CNN network then maps this raw image regions to the feature space $< \boldsymbol{f}_l(X_i^{a+}), \boldsymbol{f}_l(X_i^{a-}), \boldsymbol{f}_l(X_i^a) >$, such that the distance of the learned features of the same person is less than the distance between the images from different persons by a defined margin. Hence, the goal of the network is to minimize the following cost function for N triplet images:

$$J = \frac{1}{N}\sum_{i=1}^{N}[\left\|\boldsymbol{f}_l(X_i^a) - \boldsymbol{f}_l(X_i^{a+})\right\|_2^2 - \left\|\boldsymbol{f}_l(X_i^a) - \boldsymbol{f}_l(X_i^{a-})\right\|_2^2 + \alpha], \quad (1)$$

where α is a predefined margin which helps the network to learn more discriminative features. It is important to note that the above triplet architecture is used only in the training phase and after the network is trained, two of the three CNN channels (with shared parameters) are used for comparing two incoming images of people in testing phase. The distances between each pair of images are computed and used for ranking the pairs of query and candidate frames.

3.2. Gradient-based Attention Network (GAN)

The proposed Gradient-based Attention Network (GAN) is capable of extracting information from an image by adaptively selecting the most informative image regions and only processing the selected regions at high resolution. The whole model comprises of two blocks: the global attention network G and the local deep CNN network L. The global network consists of only two layers of convolution and is computationally efficient, whereas the local network is deeper (i.e., five convolutional layers) and is computationally more expensive, but has better performance.

We refer to the learned feature representation of the global layer and the local layer by $\boldsymbol{f}_g(I)$ and $\boldsymbol{f}_l(I)$, respectively. The attention model uses backpropagation to identify the few vectors in the global feature representation $\boldsymbol{f}_g(I)$ to which the distribution over the output of the network (i.e., $\boldsymbol{h}_g(I)$) is most sensitive. Given the input image I, we first apply the global layers on all the input regions: $\boldsymbol{f}_g = \{\boldsymbol{g}_{i,j}|(i,j) \in [1, s_1] \times [1, s_2]\}$, where s_1 and s_2 are spatial dimensions that depend on the image size and $\boldsymbol{g}_{i,j} = \boldsymbol{f}_g(I_{i,j}) \in \mathbb{R}^D$ is a feature vector associated with the input region (i, j) in I, i.e., corresponds to a specific receptive field or a patch in the input image. On top of the convolution layers in attention model, there exists a fully connected layer followed by a max pooling and a softmax layer, which consider the bottom layers' representations $\boldsymbol{f}_g(I)$ as input and output a distribution over labels, i.e., $\boldsymbol{h}_g(I)$. Next, the goal is to find an attention map. We use the entropy of the output vector $\boldsymbol{h}_g(I)$ as a measure of saliency in the following form:

$$H = \sum_{l=1}^{C} \boldsymbol{h}_g^l \log(\boldsymbol{h}_g^l), \tag{2}$$

where C is the number of class labels in the training set. In order to find the attention map we then compute the norm of the gradient of the entropy H with respect to the feature vector $\boldsymbol{g}_{i,j}$ associated with the input region (i, j) in the input image:

$$A_{i,j} = \left\| \nabla_{\boldsymbol{g}_{i,j}} H \right\|_2, \tag{3}$$

hence, the whole attention map would be $\boldsymbol{A} \in \mathbb{R}^{s_1 \times s_2}$ for the whole image. Using the attention map $\boldsymbol{A}$, we select a set of k input region positions (i, j) corresponding to the $A_{i,j}$s with the k largest values. We denote the selected set of positions by $p^a \in [1, s_1] \times [1, s_2]$ such that $\#p^a = k$. The selected regions of the input image corresponding to the selected positions are denoted by $X^a = \{x_{i,j}|(i, j) \in p^a\}$, where each $x_{i,j}$ is a patch in input image I. *Exploiting the gradient of the entropy as the saliency measure for our attention network encourages selecting the input regions which have the maximum effect on the uncertainty of the model predictions.* Note that all the elements of the attention map $\boldsymbol{A}$ can be calculated efficiently using a single pass of backpropagation. For training of the global attention network (G), we maximize the log-likelihood of the correct labels (using cross-entropy objective function).

After selecting the salient patches (X^a) within the input image, the local deep network (L) will be applied only on those patches. This leads to major saving in computational cost of the network and accuracy improvement by focusing on the informative regions of the person's image. The local deep CNN network (L) is trained on attended parts of the input image using the triplet loss introduced in Eq. 1. We denote the feature representation created by the local deep network L as $\boldsymbol{f}_l(X^a)$. In the test time, the local feature representation $\boldsymbol{f}_l(X^a)$ and the global feature representation $\boldsymbol{f}_g(I)$ can be fused to create a refined representation of the whole image. That is:

$$\boldsymbol{f}_r(x) = \begin{cases} \boldsymbol{f}_l(x_{i,j}), & \text{if } x_{i,j} \in X^a \\ \boldsymbol{f}_g(x_{i,j}), & \text{otherwise,} \end{cases} \tag{4}$$

where $(i, j) \in [1, s_1] \times [1, s_2]$. Thus, we are replacing the global features corresponding to the attended regions with the rich local features. In this way we are using all the produced features in both local and global networks and at the same time we are paying more attention to the important parts of the input image. It is worth noting that at the test time, different methods for fusion of the local and global features can be used. For instance, we can only use the local features and discard the global features. We discuss this matter further in the next section and show the evaluation results for each case.

4. EXPERIMENTS AND RESULTS

4.1. Network design

We implement our network using TensorFlow [33] deep learning framework on Intel Xeon CPU and NVIDIA TITAN X GPU. In each global attention network (see Figure 1), there are 2 convolutional layers, with 7×7 and 3×3 filter sizes, 12 and 24 filters, respectively. On the top of the two convolution layers in the global attention network there are one fully connected layer, a max pooling and a softmax layer. The global attention network is trained once for the whole network using the backpropagation algorithm. The set of selected patches X^a is composed of eight patches of size 14×14 pixels. In the local deep CNN model, there are 5 convolutional layers, each with 3×3 filter sizes, 1×1 strides, and 24 filters. We apply 2×2 pooling with 2×2 stride after the second and fourth layers. On the top of this network there is also a max pooling and a fully connected layer. We use Batch Normalization [34] and Adam [35] for training our model. We have employed the same scheme for choosing the image triplets and performing the data augmentation as in [27]. Furthermore, we have used $\alpha = 0.02$ in Eq. 1 and exponential learning rate decay for the training.

4.2. Datasets

There are several benchmark datasets for evaluation of different person re-identification algorithms. In this paper we use CUHK01 [5] and CUHK03 [6] which are two of the largest benchmark datasets suitable for training the deep convolutional network. The CUHK01 dataset contains 971 persons captured from two camera views in a campus environment. Camera view A captures frontal or back views of a person while camera B captures the person's profile views. Each person has four images with two from each camera. The CUHK03 dataset contains $13,164$ images of $1,360$ identities. All pedestrians are captured by six cameras, and each person's image is only taken from two camera views. It consists of manually cropped person images as well as images that are automatically detected for simulating more realistic experiment situation. In our experiments we used the cropped person images.

4.3. Evaluation Metric and Results

We adopt the widely used Cumulative Match Curve (CMC) metric for quantitative evaluations. For each dataset, we randomly select 100 images of the persons for testing, and the remaining for training. For datasets with two cameras, we randomly select one image of a person from camera A as a query image and one image of the same person from camera B as a gallery image. For each image in the query set, we first compute the distance between the query image and all the gallery images using the L_2 distance and then return the top n nearest images in the gallery set. If the returned list contains an image featuring the same person as that in the query image at k-th position, then this query is considered as success of rank k. We repeat the procedure 10 times, and use the average rate as the evaluation result.

As it was mentioned in Sec. 3, at the test time, we can use different combinations of the features that are learned by the global attention network (i.e., $\boldsymbol{f}_g(I)$) and local deep network (i.e., $\boldsymbol{f}_l(X^a)$). We performed the experiment in two scenarios; one is using only the local deep network features $\boldsymbol{f}_l(X^a)$ for calculating the distance between the test image pairs. We denote this results by **GAN-L** in Tables 1 and 2. Another scenario is using the fusion of the local and global features (i.e., $\boldsymbol{f}_r(x)$ in Eq. 4) at test time, denoted by **GAN** in the following tables. We compare the re-identification results with state-of-the-art methods: Imp-Reid [28], FPNN [6], SDALF [36], eSDC [15], KISSME [26] and Partb-ReId [27]. Results are illustrated in Table 1 and 2 for CUHK01 and CUHK03 datasets, respectively.

We observe from Tables 1 and 2 that the proposed GAN method outperforms other state-of-the-art methods in most cases for CUHK01 and CUHK03. Table 1 shows that the performance of the method in [27] is comparable to the proposed method in some cases. However, the GAN is computationally more efficient since it first detects the most discriminative parts of the input image and then it applies the local deep convolution networks on those few patches. However, in [27], the global and local convolution layers are applied on all parts without considering the saliency of the image regions. As it is shown in Tables 1 and 2, the GAN network which exploits the refined feature presentation obtained from fusion of local and global features achieves better performance compared to GAN-L which only employs the fine local features corresponding to the attended regions of the image. Furthermore, Table 2 shows that GAN achieves the best re-identification accuracy for rank 1 and rank 5 and second best accuracy for rank 10 and rank 20 on CUHK03 dataset. Moreover, in our experiments we observed that when the network chooses to attend to the upper body part of the persons' images, the network achieves higher matching rating. In fact, this observation confirms our daily experience in re-identifying people where we usually pay attention to the face and upper body part of the person.

Table 1: Comparison of performance of the proposed GAN to the state-of-the-art on CUHK01 dataset

Method	Rank1	Rank5	Rank10	Rank20
FPNN [6]	22.87	58.20	73.46	86.31
SDALF [36]	9.90	41.21	56.00	66.37
eSDC [15]	22.84	43.89	57.67	69.84
KISSME [26]	29.40	57.67	72.42	86.07
Partb-reid [27]	53.7	84.3	**91.0**	96.3
GAN-L	54.6	83.6	89.4	90.2
GAN	**64.2**	**86.4**	90.6	**96.9**

Table 2: Comparison of performance of the proposed GAN to the state-of-the-art on CUHK03 dataset

Method	Rank1	Rank5	Rank10	Rank20
Imp-reid [28]	54.74	86.50	**93.88**	**98.10**
FPNN [6]	20.65	51.50	66.50	80.00
SDALF [36]	5.60	23.45	36.09	51.96
eSDC [15]	8.76	24.07	38.28	53.44
KISSME [26]	14.17	48.54	52.57	70.53
GAN-L	60.5	82.2	88.8	91.5
GAN	**61.2**	**89.1**	91.3	93.9

5. CONCLUSION

In this paper we introduced a CNN-based attention mechanism for person re-identification task and we showed how paying attention to important parts of the person's image while still considering the whole image information, is beneficial for person re-identification. Furthermore, thanks to the computational efficiency resulting from the attention architecture, we would be able to train deeper neural networks in order to obtain higher accuracy in the re-identification task.

6. REFERENCES

[1] Kelvin Xu, J Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention.," in *ICML*, 2015, vol. 14, pp. 77–81.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu, "Multiple object recognition with visual attention," *arXiv preprint arXiv:1412.7755*, 2014.

[4] Ronald J Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[5] Wei Li and Xiaogang Wang, "Locally aligned feature transforms across views," 2013, pp. 3594–3601.

[6] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *(CVPR)*, 2014, pp. 152–159.

[7] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li, "Person re-identification by local maximal occurrence representation and metric learning," in *(CVPR)*, 2015, pp. 2197–2206.

[8] Alireza Rahimpour, Ali Taalimi, Jiajia Luo, and Hairong Qi, "Distributed object recognition in smart camera networks," in *IEEE International Conference on Image Processing, Phoenix, Arizona, USA*. IEEE, 2016.

[9] Behnam Babagholami-Mohamadabadi, Amin Jourabloo, Ali Zarghami, and Mahdieh Soleymani Baghshah, "Supervised dictionary learning using distance dependent indian buffet process," in *Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on*. IEEE, 2013, pp. 1–6.

[10] Sateesh Pedagadi, James Orwell, Sergio Velastin, and Boghos Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *(CVPR)*, 2013, pp. 3318–3325.

[11] Behnam Babagholami-Mohamadabadi, Seyed Mahdi Roostaiyan, Ali Zarghami, and Mahdieh Soleymani Baghshah, "Multi-modal distance metric learning: Abayesian non-parametric approach," in *European Conference on Computer Vision*. Springer, 2014, pp. 63–77.

[12] Chi Su, Fan Yang, Shiliang Zhang, Qi Tian, Larry S Davis, and Wen Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3739–3747.

[13] Alireza Rahimpour and Abbas Nasiraei M., "Eye tracking by image processing for helping disabled people," *Iranian Journal of Biomedical Engineering (IJBME)*, vol. 6, no. 3, pp. 195–205, 2012.

[14] Rahul Rama Varior, Gang Wang, Jiwen Lu, and Ting Liu, "Learning invariant color features for person re-identification," in *IEEE Transaction on Image processing*. 2016, IEEE.

[15] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Unsupervised salience learning for person re-identification," in *(CVPR)*, 2013, pp. 3586–3593.

[16] Mostafa Rahmani and George Atia, "Spatial random sampling: A structure-preserving data sketching tool," *arXiv preprint arXiv:1705.03566*, 2017.

[17] Shengcai Liao, Guoying Zhao, Vili Kellokumpu, Matti Pietikäinen, and Stan Z Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 1301–1306.

[18] Alireza Rahimpour, Ali Taalimi, and Hairong Qi, "Feature encoding in band-limited distributed surveillance systems," in *ICASSP 2017-IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2017.

[19] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[20] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, "Person re-identification by salience matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2528–2535.

[21] Mostafa Rahmani and George Atia, "Coherence pursuit: Fast, simple, and robust principal component analysis," *arXiv preprint arXiv:1609.04789*, 2016.

[22] Alireza Rahimpour, Hairong Qi, David Fugate, and Teja Kuruganti, "Non-intrusive energy disaggregation using non-negative matrix factorization with sum-to-k constraint," *IEEE Transactions on Power Systems*, 2017.

[23] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[24] Seyyed Ali Davari, Sheng Hu, and Dibyendu Mukherjee, "Calibration-free quantitative analysis of elemental ratios in intermetallic nanoalloys and nanocomposites using laser induced breakdown spectroscopy (libs)," *Talanta*, vol. 164, pp. 330–340, 2017.

[25] Alireza Rahimpour, Hairong Qi, David Fugate, and Teja Kuruganti, "Non-intrusive load monitoring of hvac components using signal unmixing," in *Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on*. IEEE, 2015, pp. 1012–1016.

[26] Martin Köstinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2288–2295.

[27] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *(CVPR)*, 2016, pp. 1335–1344.

[28] Ejaz Ahmed, Michael Jones, and Tim K Marks, "An improved deep learning architecture for person re-identification," in *(CVPR)*, 2015, pp. 3908–3916.

[29] Jane Bromley, James W Bentz, Léon Bottou, Isabelle Guyon, Yann LeCun, Cliff Moore, Eduard Säckinger, and Roopak Shah, "Signature verification using a siamese time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, no. 04, pp. 669–688, 1993.

[30] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[31] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan, "End-to-end comparative attention networks for person re-identification," *arXiv preprint arXiv:1606.04404*, 2016.

[32] Amjad Almahairi, Nicolas Ballas, Tim Cooijmans, Yin Zheng, Hugo Larochelle, and Aaron Courville, "Dynamic capacity networks," *arXiv preprint arXiv:1511.07838*, 2015.

[33] Martín Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, Software available from tensorflow.org.

[34] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[35] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2360–2367.