End-to-End Correspondence and Relationship Learning of Mid-Level Deep Features for Person Re-Identification

Shan Lin and Chang-Tsun Li

***Abstract*— In this paper, a unified deep convolutional architecture is proposed to address the problems in the person re-identification task. The proposed method adaptively learns the discriminative deep mid-level features of a person and constructs the correspondence features between an image pair in a data-driven manner. The previous Siamese structure deep learning approaches focus only on pair-wise matching between features. In our method, we consider the latent relationship between mid-level features and propose a network structure to automatically construct the correspondence features from all input features without a pre-defined matching function. The experimental results on three benchmarks VIPeR, CUHK01 and CUHK03 show that our unified approach improves over the previous state-of-the-art methods.**

I. INTRODUCTION

Person re-identification (re-ID) addresses the problem of matching different persons across disjoint camera views[1]. It has potential applications in video surveillance and multimedia forensics such as pedestrian retrieval, cross-camera tracking, and activity and event detection. A typical person re-identification system can be separated into three different modules: people detection, people tracking, and people retrieval. The first two modules are considered as independent computer vision problems which have already been researched over decades. Therefore, the main task of person re-identification focuses on the person retrieval module by constructing robust features for distinctive appearance representation of people and developing better matching strategies.

There are many people re-identification models developed by exploiting low-level features such as colour [2], texture, spatial structure [3], *etc*. However, these low-level visual features are not robust to variations in illumination, viewpoint, misalignment, *etc*. In human perception, different people can be easily recognised by their mid-level visual features such as long hair, blue shirts, green handbag, *etc*. These attributes can represent the mid-level semantics of a person and are more robust to misalignment and variations comparing to low-level local features. However, manual annotation of these mid-level semantics features is very expensive. As a result, it is difficult to acquire enough training data with a large set of labelled human attributes.

Our proposed method uses an alternative approach to obtain the mid-level features. In recent years, deep convolutional neural networks (dCNN) architectures show a significant improvement in performance when solving computer vision tasks. There are also many studies analysing the features obtained by dCNN. As the features from dCNN are structured in a hierarchical nature, the lower layer behaves similarly to low-level local feature extractors such as edge or colour filters. At higher layers, the features start showing significant variation and become more class-specific [4]. In our proposed method, We made two assumptions:

1) The features maps from higher-level layers of dCNN are good representations of the mid-level features for human appearance.
2) Using an ImageNet-1K [5] pre-trained model as the initial state of the convolutional neural network feature extractor is beneficial.

If we have a good matching function to find the correspondence between features, the feature extractor can be trained to capture the most distinctive features for person representation. Therefore, the remaining task is to determine the proper matching strategy. Many existing approaches focus on constructing the correspondence distributions between each pair of the same feature map from the probe and gallery images. In our point of view, the mid-level feature correspondences should not be limited to pair-wise feature map

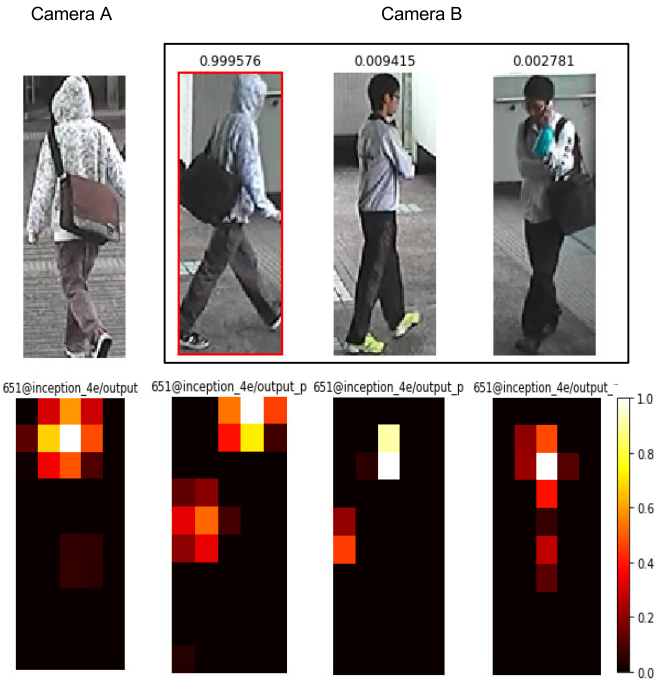

Fig. 1. This figure shows our predicted results in the CUHK01 dataset. The ground-truth images are marked by the red bounding boxes. The second line shows one of the mid-level features for each person obtained in our network. The white hoodie hat is one of the distinctive mid-level features when re-identifying the person in the probe image

matching. The potential relationship between these mid-level features should also be taken into consideration. We propose a new strategy for establishing feature correspondence by considering different combinations of mid-level features. In our proposed network, each correspondence feature is not limited to the correlation between single feature map from two images, but from the multiple features maps of two images. Furthermore, our network can also capture the relationship between these correspondence features in a data-driven manner.

In this paper, we propose an end-to-end deep learning framework for assigning a similarity score to each pair of images of pedestrians. One example of our system prediction result is shown in Fig 1. By using the Inception network [6] as the mid-level feature extractor, the proposed method can adaptively discover the intra-personal and inter-personal relationships of the mid-level deep features. The similarity score is calculated by analysing the relationship between the correspondence features. In addition, these latent relationships between mid-level features are considered and learned through a data-driven approach for generalising the representations of people. It shows greater robustness in the cross-dataset scenarios. Furthermore, as the parameters are initialised from the pre-trained ImageNet model, the training process of our network can be considered as a fine-tuning process for regularising the deep mid-level features from object classification to similarity matching. As a result, it improves the discriminative power of these deep features.

II. Related Work

Typical person retrieval process includes two components: a method for extracting features from input images, and a similarity metric for comparing those features across images. The main objective of searching better feature representations is to find features which are relatively invariant to lighting condition, human poses and camera viewpoints. The early approaches relied on the handcrafted features including HSV colour histogram [3], LBP and Garbo features [7], SIFT [8], *etc*. With all these features, many similarity metrics are also proposed such as Mahalanobis metric learning [9], LADF [10], saliency weighted distances [11], *etc*.

In recent years, due to the promising performance of deep learning, many researchers began to use deep learning to obtain the visual features and distance metrics for person re-identification [12], [13], [14]. The deep metric learning approach [15] partitions the input image into three overlapping horizontal parts, and these parts go through two convolutional layers plus a fully connected layer which fuses them and outputs a vector for this image. The similarity of the two output vectors is computed using the cosine distance. The FPNN architecture [12] is different in that a patch matching layer is added, which multiplies the convolution responses of two images in different horizontal stripes. The ImprovedReID [13] improved the FPNN model by computing the cross-input neighbourhood difference features, which compares the features from one input image to features in neighbouring locations of the other image. However, such matching strategies may either be of low computing efficiency or have limitations due to the lack of the spatial structure information.

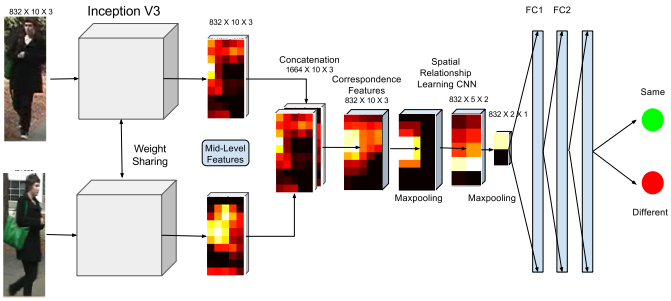

Fig. 2. Proposed Network Architecture

III. Proposed Approach

The framework of our architecture is shown in Fig 2. The network can be divided into three components:

1) **The mid-level image representations of each image are extracted from Inception convolutional layer**: *(Inception_4e/output)*.
2) **The mid-level feature correlations between two images and the correspondence relationship between related features are learned by using multi-layers convolutional neural networks.**
3) **The metric network of three fully connected layers is used for computing the similarity score.**

A. Mid-level Feature Extraction

The ImageNet-1K pre-trained Inception model [16] is used as the basic deep neural network architecture for creating the Siamese network structure in our architecture. Two networks serve as feature extractors for obtaining the mid-level feature maps of each input image. In order for the features to be comparable, the weights of all convolutional layers for the feature extraction process are shared. The initial weights are generated by training on the ImageNet-1K dataset and then transferred to our network. The ImageNet pre-trained model serves as a good starting point for the later network training and fine-tuning.

As our training objective differs from ImageNet classification task, the input image shape is not restricted to the 256×256 image shape from the ImageNet. In our architecture, we decide to normalised all the input images to 160×80 which is similar to the height-width ratio of images in majority person re-identification datasets and generates less distortion to the original images.

In our architecture, we decided to use the lower level "*Inception_4e/output*" layer instead of the last Inception convolutional layer (*Inception_5b/output layer*) as our mid-level feature outputs. There are two reasons:

- With the 160×80 input shape, the last convolutional layer outputs will be 5×2 which loses too much spatial structure information. The feature maps from

the "*Inception_4e/output*" are relatively larger with the 10×5 in shape.
- The last convolutional layer of the deep learning model produces high-level features. In the person re-identification situation, mid-level features are more suitable for the task. Therefore, we use the feature maps from a lower level convolutional layer to represent the mid-level features.

B. Correspondence Features Learning

Given a probe in camera A and a gallery image in camera B, each image is represented by 832 feature maps after the mid-level feature extraction process, detailed in the section III-A above. Let X_i^A and X_i^B represent the ith mid-level feature map ($1 \leq i \leq 832$) extracted from two input images. The similarity between the people in the probe and gallery can be learned by analysing the correspondence relationship between X_i^A and X_i^B of the image pair. The previous approaches focus on learning the correspondence features by calculating pair-wise matching probabilities. For example, the first feature map from probe and gallery images, X_0^A and X_0^B are divided into patches. The correspondence feature of the first feature map is obtained by dense patch matching [17] or local searching in the neighbourhood of the given location [13]. However, with these approaches, each correspondence feature is obtained from only a pair of respective feature maps like [X_0^A, X_0^B] or [X_1^A, X_1^B]. They assume the extracted features maps are independent and fail to address the possible latent relationship among different feature maps. For example, feature maps 1, 3 and 6 can be grouped together to give a better correspondence feature: [$X_{1,3,6}^A$, $X_{1,3,6}^B$].

In our proposed method shown in Fig 2, our correspondence features obtained by using a convolutional layer:

$$C = f_*([X^A, X^B], \Theta)$$

where f_* denotes the convolution operation. $[X^A, X^B]$ is the concatenation of two mid-level feature maps with shape $1664 \times 3 \times 10 \times 5$. With kernel size 3, padding 1 and stride 1, the output feature maps can maintain the shape of $3 \times 10 \times 5$. The number of output feature maps is set to be the same as the mid-level feature maps, to represent the 832 correspondence features. As the convolution operation is performed on all mid-level feature maps, each output feature can be considered as one possible feature correlation between all feature maps of each image pair. In our proposed method, the correspondence features are not limited to the specific pair of feature maps, but learned from the combinations of many different feature maps. All the weights for combination and convolutional filters are automatically learned in a data-driven manner.

C. Spatial Relationship Learning

One of the biggest problems in person re-identification is misalignment. In order to learn the spatial relationship between all these correspondence features, another convolutional layer is introduced. The input and output shapes are the same ($832 \times 3 \times 10 \times 5$) with kernel size of 3, padding 1 and stride 1. To deal with viewpoint variation and misalignment, the max-pooling layer is used to further reduce the spatial size of the representation and align to the correspondence features to large regions. The convolutional layer after max-pooling is used to learn the relationship from a different scale.

D. The Metric Network

Inspired by the MatchNet [18], our similarity metric between features is modelled by using three fully-connected layers with the ReLU non-linearity activation function. The output of the last fully-connected layer will be two values in the range of [0,1]. They can be interpreted as the probability whether the two input images are capturing the same person or not. Besides, we also add a dropout layer after the first and second fully-connected layers to reduce the over-fitting problem and obtain better generalisation ability.

E. Loss Function

Our network is trained and optimised by minimising the cross-entropy error of the output labels using stochastic gradient descent:

$$E = -\frac{1}{N}\sum_{n=1}^{N}[y_n \cdot log(\hat{y}_n) + (1 - y_n) \cdot log(1 - \hat{y}_n)]$$

N refers to the number of image pairs used in a mini-batch during training. Here y_n is the ground truth of image pair x_n. $y_n = 1$ indicates the image pair is the same person and $y_n = 0$ means negative matching. $\hat{y}_n$ is the Softmax activation computed based on the output value from the two nodes in the last fully-connected layer $v_0(x_n)$ and $v_1(x_n)$:

$$\hat{y}_n = \frac{e^{v_1(x_n)}}{e^{v_1(x_n)} + e^{v_0(x_n)}}$$

In summary, our proposed method can adaptively obtain the mid-level features, automatically construct the correspondence features with their relationship and finally learn the similarity metric in a data-driven manner. Comparing to previous one-to-one feature map matching approaches, we consider the latent relationship between features when learning the correspondence features.

IV. Model Training

A. Dataset

Three publicly available datasets are used for evaluation: VIPeR [19], CUHK01[20] and CUHK03[12]. The most tested benchmark is the VIPeR dataset. It contains 632 identities and two images for each identity. The CUHK01 dataset was also captured from two camera views. It has 971 persons, and each person has two images from camera A, and the other two from camera B. Camera A takes a frontal view and Camera B, the side view. The CUHK03 dataset contains 13,164 images of 1,360 pedestrians, captured by six surveillance cameras. Each identity is observed by two disjoint camera views. On average, there are 4.8 images per identity from each view. The statistics of these datasets are summarised in Table I below.

TABLE I

STATISTICS OF EACH DATASET

Dataset	#ID	#Image	#Camera	label
VIPeR	632	1264	2	hand
CUHK01	971	3884	2	hand
CUHK03	1360	13164	2	hand/DPM

B. Training

In the training process, the training image pairs are divided into mini-batches of size 100. Therefore, the total number of batches are over one hundred thousand. The stochastic gradient descent is used as the optimisation method for minimising the cross-entropy error. The learning rate is 0.01 with polynomial decay. The momentum is set to 0.9 with the weight decay of 0.0002.

C. Compensation for Training Data Imbalance

For each person in the datasets, there are only a few positive matching images with a huge amount of negative matching images. Therefore, during the training process, the number of positive images pairs will be much less than negative pairs which can lead to data imbalance and over-fitting. To reduce the potential over-fitting, we also implemented two commonly used preprocessing methods [13]:

1) ***Data Augmentation****:* The original training images are reshaped under random 2D affine transformations around the image center to obtain extra five augmented images. The smaller dataset such as VIPeR will be further augmented by horizontally reflected. This process will not only reduce the data imbalance problem but also generate more training samples.

2) ***Hard Negative Mining****:* Data augmentation increases the number of positive pairs, but the training dataset is still imbalanced with many more negatives than positives. If we trained the network with this imbalanced dataset, it may learn to predict every pairs as negative. Therefore, we randomly down-sample the negative sets to get just twice as many negatives as positives (after augmentation), then train the network. The converged model obtained is not optimal since it has not seen all possible negatives. We use the current model to classify all of the negative pairs, and choose the top ranked ones which our network performs the worst for retraining our network.

D. Visualisation of Mid-level Feature

Fig 3 gives a visualisation of one mid-level feature learned after the training process. They are the highest weighted feature map from the "*Inception_4e/output*" layer when extracting the mid-level features from two images of the same person. The region with very light colour means high activation values. In this case, the most activated region is highlighted around her green handbag. From this experiment, we realised that many mid-level feature maps obtained from our proposed network have semantic meanings which can be very useful for later feature correspondence learning. As a result, it proves that our network can successfully learn good mid-level features for human representation.

Fig. 3. #171 activation feature map from inception_4e/output detecting the handbag

V. EXPERIMENTAL RESULTS

In this section, the performance of our model is compared with several methods developed in recent years such as KISSME [9], SalMatch [21], FPNN [12], ImprovedReID [13], LMNN [22], DML [15] and XQDA+LOMO [23]. We adopt the widely used cumulative match curve (CMC) approach for quantitative evaluation

A. Experiments on CUHK01

The CUHK01 dataset contains 3884 images of 971 identities from two different cameras. Previous state-of-the-art approaches normally have two different settings for this dataset: 100 test IDs and 486 test IDs [21], [23]. As the deep learning approaches require a large dataset for training, we did not perform the 486 test IDs experiment. In our experiment, we only focus on 100 randomly selected identities for testing. The remaining identities are used for training. Table II is the comparison of our proposed method with the recent state-of-art results. Our method outperforms the ImprovedReID in this setting by a large margin. The CMC curves of all these methods are shown in Fig 4.

TABLE II

COMPARISON WITH THE STATE-OF-ART RESULTS REPORTED ON THE CUHK01 DATASET

Methods	Rank 1	Rank 5	Rank 10
SDALF	9.9	41.2	56.90
LMNN	21.2	48.5	62.9
FPNN	27.9	48.5	63.0
KISSME	29.4	60.18	74.4
SalMatch	28.5	45.0	55.0
XQDA+LOMO	63.2	83.9	90.0
ImprovedReID	65.0	89.0	94.0
Proposed	81.2	95.8	97.4

B. Experiments on CUHK03

The CUHK03 dataset contains 13164 images of 1360 identities from six different cameras. This dataset has two different pedestrians datasets. One is manually labelled while the other is extracted from Deformable Parts Model (DPM) human detector [24]. Our model is tested based on the labelled dataset. Table III is the comparison of our methods

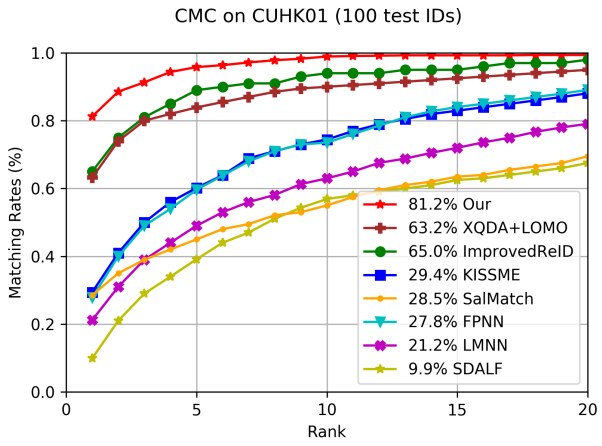

Fig. 4. CMC curves on the CUHK01 dataset

with the recent state-of-art results. Overall deep learning approaches such as FPNN and ImprovedReID show better results on large datasets when compared to many traditional handcrafted features and learning metrics approaches. Our model still outperforms the ImprovedReID from 55% to 72% in rank-1 accuracy and yields over 90% rank-5 accuracy. The detail CMC performance comparison with other models are shown in Fig 5.

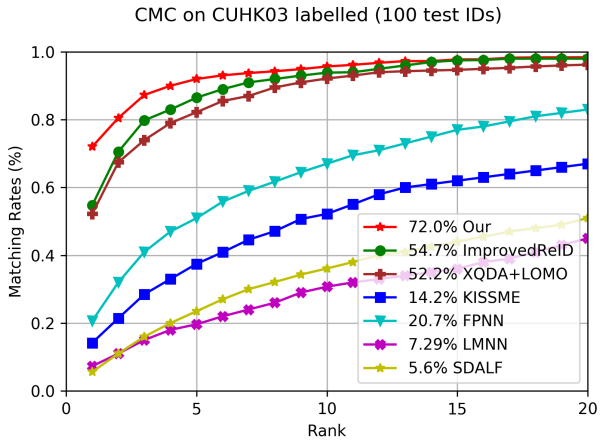

Fig. 5. CMC curves on the CUHK03 labelled dataset

C. Experiments on VIPeR

As the VIPeR dataset is very small, the dataset alone cannot provide enough training data for deep learning methods to properly coverage. Therefore, the ImprovedReID and our proposed method have to pre-train on the combination of the CUHK03 and CUHK01 datasets, then fine-tuned on VIPeR training data. The rest of the traditional approaches such as KISSME and XQDA+LOMO follow the commonly applied 50% training and 10 fold cross-validation evaluation. Table IV below illustrates the overall performance of our model. It outperforms very well to the state-of-art methods even with a small fine-tuned training sample.

TABLE III
COMPARISON WITH THE STATE-OF-ART RESULTS REPORTED ON THE CUHK03 LABELLED DATASET

Methods	Rank 1	Rank 5	Rank 10
SDALF	5.6	23.5	36.1
LMNN	7.3	19.6	30.7
FPNN	20.6	50.9	67.1
KISSME	14.7	37.3	52.2
XQDA+LOMO	52.2	82.2	93.9
ImprovedReID	54.7	86.5	93.9
Proposed	72.0	92.7	95.7

TABLE IV
COMPARISON WITH THE STATE-OF-ART RESULTS REPORTED ON THE VIPER DATASET

Methods	Rank 1	Rank 5	Rank 10
KISSME	19.6	48.0	62.2
SalMatch	30.2	52.0	65.5
LMNN+LOMO	29.4	59.8	73.5
KISSME+LOMO	34.8	60.4	77.2
XQDA+LOMO	40.0	68.1	80.5
DML	28.2	59.3	73.5
ImprovedReID	34.8	63.6	75.6
Proposed	42.5	71.4	80.6

D. Cross-dataset Evaluations

As our model can adaptively obtain the great correspondence of mid-level features and learn the relationship between them, we believe that it should have the ability to generalise to distinctive features and a similarity metric network for person re-identification tasks in the cross-dataset scenario. In the experiment, our model after training on the full CUHK03 dataset can achieve 64.2% rank-1 accuracy when tested on the full CUHK01 dataset (similar performance to XQDA+LOMO model on the CUHK01 dataset) and 14.5% rank-1 accuracy on the VIPeR (similar performance to KISSME model on the VIPeR dataset)

VI. CONCLUSION

We proposed an effective end-to-end deep learning approach for the person re-identification task in this paper. Unlike previous deep learning approaches, our model considers the possible latent relationship between mid-level features when generating the feature correspondences. Our feature correspondences give a more robust representation of an image pair. Benefiting from the latent mid-level feature correspondences learning, our proposed method obtains superior performance compared to many state-of-the-art approaches on three benchmark datasets, VIPeR, CUHK01 and CUHK03. In addition, as an end-to-end network, our network can simultaneously learn the deep mid-level features, feature correspondences and correlation relationship as well as a metric learning network for solving the misalignment and viewpoint variation problems in person re-identification.

Acknowledgement: This work is supported by EU Horizon 2020 project, entitled Computer Vision Enable Multimedia Forensics and People Identification (acronym:IDENTITY, Project ID:690907)

REFERENCES

[1] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable Person Re-identification: A Benchmark," in *International Conference on Computer Vision (ICCV)*, vol. 11-18-Dece, pp. 1116–1124, IEEE, 12 2015.

[2] C. Madden, E. D. Cheng, and M. Piccardi, "Tracking People across Disjoint Camera Views by an Illumination-Tolerant Appearance Representation," *Machine Vision and Applications*, vol. 18, pp. 233–247, 5 2007.

[3] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-Driven Accumulation of Local Features for Human Characterization and Re-identificatio," *Computer Vision and Image Understanding*, vol. 117, pp. 130–144, 2 2013.

[4] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *European Conference on Computer Vision (ECCV)*, vol. 8689, pp. 818–833, 2014.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, pp. 1–9, 2012.

[6] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, IEEE, 6 2015.

[7] W. Li and X. Wang, "Locally Aligned Feature Transforms across Views," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3594–3601, IEEE, 6 2013.

[8] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised Salience Learning for Person Re-identification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3586–3593, IEEE, 6 2013.

[9] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large Scale Metric Learning from Equivalence Constraints," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2288–2295, IEEE, 6 2012.

[10] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning Locally-Adaptive Decision Functions for Person Verification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3610–3617, IEEE, 6 2013.

[11] N. Martinel and C. Micheloni, "Saliency Weighted Features for Person Re-Identification," in *ECCV Workshop on Visual Surveillance and Re-identificatio* (L. Agapito, M. M. Bronstein, and C. Rother, eds.), vol. 8927 of *Lecture Notes in Computer Science*, (Cham), pp. 0–17, Springer International Publishing, 2014.

[12] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 152–159, IEEE, 6 2014.

[13] E. Ahmed, M. Jones, and T. K. Marks, "An Improved Deep Learning Architecture for Person Re-Identification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3908–3916, IEEE, 6 2015.

[14] S. Paisitkriangkrai, C. Shen, and A. van den Hengel, "Learning to Rank in Person Re-identification with Metric Ensembles," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 07-12-June, pp. 1846–1855, IEEE, 6 2015.

[15] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep Metric Learning for Person Re-identification," in *International Conference on Pattern Recognition (ICPR)*, no. 1, pp. 34–39, IEEE, 8 2014.

[16] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

[17] Y. Shen, W. Lin, J. Yan, M. Xu, J. Wu, and J. Wang, "Person re-identification with correspondence structure learning," *International Conference on Computer Vision (ICCV)*, vol. 11-18-Dece, pp. 3200–3208, 2015.

[18] Xufeng Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "MatchNet: Unifying Feature and Metric Learning for Patch-based Matching," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3279–3286, IEEE, 6 2015.

[19] D. Gray, S. Brennan, and H. Tao, "Evaluating Appearance Models for Recognition, Reacquisition, and Tracking," *International Workshop on Performance Evaluation for Tracking and Surveillance (PETS),*, vol. 3, pp. 41–47, 2007.

[20] W. Li, R. Zhao, and X. Wang, "Human Reidentification with Transferred Metric Learning," in *Asian Conference on Computer Vision (ACCV)* (K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, eds.), no. PART 1, pp. 31–44, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

[21] R. Zhao, W. Ouyang, and X. Wang, "Person Re-identification by Salience Matching," in *International Conference on Computer Vision (ICCV)*, pp. 2528–2535, IEEE, 12 2013.

[22] K. Q. Weinberger and L. K. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 7 2009.

[23] S. Liao, Y. Hu, Xiangyu Zhu, S. Z. Li, X. Zhu, and S. Z. Li, "Person Re-identification by Local Maximal Occurrence Representation and Metric Learning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 07-12-June, pp. 2197–2206, IEEE, 6 2015.

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 9 2010.