Unconstrained Face Verification and Open-World Person Re-identification via Densely-connected Convolution Neural Network

Donghwuy Ko, Jongmin Yu, Ahmad Muqeem Sheri and Moongu Jeon
Gwangju Institute of Science and Technology 123, Cheomdangwagi-ro, Buk-gu, Gwangju, South Korea

Keywords: Face Verification, Person Re-Identification, Unconstrained Condition, Metric Learning.

Abstract: Although various methods based on the hand-crafted features and deep learning methods have been developed for various applications in the past few years, distinguishing untrained identities in testing phase still remains a challenging task. To overcome these difficulties, we propose a novel representation learning approach to unconstrained face verification and open-world person re-identification tasks. Our approach aims to reinforce the discriminative power of learned features by assigning the weight to each training sample. We demonstrate the efficiency of the proposed method by testing on datasets which are publicly available. The experimental results for both face verification and person re-identification tasks show that its performance is comparable to state-of-the-art methods based on hand-crafted feature and general convolutional neural network.

1 INTRODUCTION

Image recognition has been studied extensively in computer vision which has resulted in the development of various methods based on the hand-crafted features as well as deep learning methods. Recently, the introduction of representation learning methods based on the deep convolutional neural network (DCNN) provided automatic feature representation and have shown outstanding accuracies. Face verification and person re-identification have also been extensively studied in computer vision studies. Face verification is a binary classification task to classify whether two input images share the same identity, while person re-identification task is identification over dataset where images of each identity are in different environment variables, such as different illumination and angle of view. To be more specific, person re-identification is the task of identifying test images from a gallery where test images and training images are from different cameras. For both tasks, representation learning methods based on DCNN showed outstanding accuracies compared to hand-crafted methods. However, in case of person re-identification, it is a still difficult problem to handle various images distorted by their extrinsic factors such as illumination and camera angles. Even though recently proposed methods based on deep learning can cover these various images, these methods only focus on feature which can be extracted using training samples. In case of face verification, it is challenging to perform face verification on unconstrained condition, where identities of test dataset and training dataset are mutually exclusive.

To overcome these difficulties over person re-identification and face verification, it is essential to develop a method that works precisely in both tasks under unconstrained condition and various environments. Both face verification and person re-identification tasks under the unconstrained condition and various environments are certainly challenging issues. To achieve high accuracy in these conditions, it is necessary to extract discriminative features. Since identities in training dataset and test dataset are different, it is obvious that generalized features would lead to better performance. Especially in person re-identification, it is important to minimize the effect of noises caused by its different environmental condition.

Minimization of noise and classification using extracted features can be interpreted as reducing intra-class variance and increasing inter-class variance, respectively. Inter-class variance is variance among centroid of different identities, whereas intra-class variance is variance among images that share the same identity. In other words, intra-class variance determines the size of clusters, and inter-class variance determines the size of the overlapping area among them. To make outstanding networks that can extract discriminative representation, both inter-class variance and intra-class variance should be considered.

Ko, D., Yu, J., Sheri, A. and Jeon, M.
Unconstrained Face Verification and Open-World Person Re-identification via Densely-connected Convolution Neural Network.
DOI: 10.5220/0007381104430449
In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 443-449
ISBN: 978-989-758-354-4

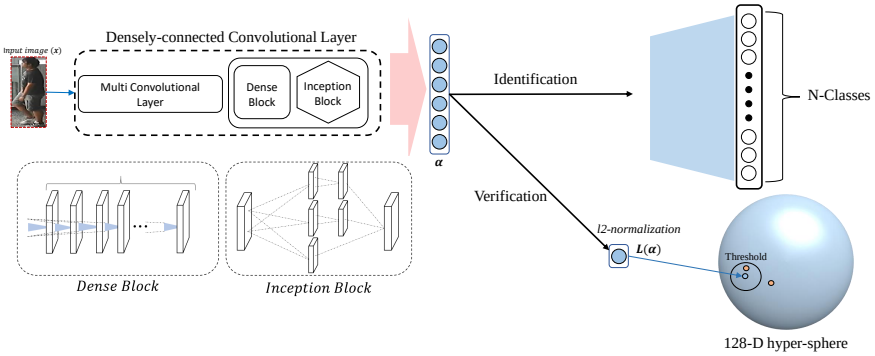

Figure 1: Overall architecture of proposed network. α represents the output of encoder. In case of verification, latent features are projected into 128-dimension hypersphere by performing *L*2-normalization. Other projections which have smaller distance than given threshold is interpreted as same identity.

The main contribution of this paper is a proposal of a simple and novel way to handle both inter-class variance and intra-class variance. We applied a method called loss-weighted gradient where elements with a higher loss in a mini-batch are weighed more to overcome difficulties of training hard samples. The network structure in this paper is inspired by Siamese Network (Bromley et al., 1994), DenseNet (Huang et al., 2016), and Inception model (Szegedy et al., 2015). We employed the densely connected convolutional neural network and inception model. Loss-weighted gradient is applied to identification loss to weight the gradient computed from hard samples than easy samples.

2 PROPOSED APPROACH

The proposed network consists of three parts: feature extraction, identification, and verification. In feature extraction, 128 dimension latent features of given images are extracted. Extracted features are then processed by a softmax layer for identification. Latent features are also projected on 128 dimension hypersphere using *L*2 normalization and passed into the verification process. These processes aim to optimize inter-class variance and intra-class variance.

2.1 Structural Details

The network structure of proposed model is inspired by DenseNet (Huang et al., 2016) and Inception (Szegedy et al., 2015). Low-level features are extracted through six conventional convolutional layers and one max-pooling layer. Eight DenseNet blocks and two inception blocks are applied subsequently. From proposed network, latent features α are encoded. In the case of identification, softmax classification is performed using α as input. In case of verification task, unlike other methods (Bromley et al., 1994) which uses latent features itself for verification, we reform latent features using L2-normalization. L2-normalized features described by:

$$L(\alpha) = \| \alpha \|_2 = \frac{\alpha}{\max(\| \alpha \|, \varepsilon)} \quad (1)$$

where ε is a parameter that prevents the output value from diverging to ∞. $L(\alpha)$ can be understood as the projected point of α onto the surface of 128-dimensional hypersphere. In our method, the distance between two L2-normalized features is interpreted as dissimilarity. When the distance between two features is close, it implies that two image have high similarity. Figure 1 illustrates the overall architecture of the proposed network.

2.2 Construction of Training Dataset

In many applications based on deep learning (Yin et al., 2011; Sun et al., 2013; Zhu et al., 2014; Sun et al., 2014), it is essential to train model with a lot of training data. In case of Person re-identification, we merge 4 publicly available datasets: CUHK-02 dataset (Li and Wang, 2013), CUHK-03 dataset (Li et al., 2014), CAVIAR4REID dataset (Cheng et al., 2011), and iLIDs-VID dataset(Wang et al., 2014), and applied the data augmentation. We applied the simple linear transformation to artificially enlarge the original dataset. Rotation, horizontal flipping, cropping, and blurring were employed to transform the original image dataset. Specifically, because the CUHK-02 dataset contains the images of the CUHK-01 dataset used in evaluation, we manually remove the duplicate images. Consequently, the training dataset consists of 217,942 images with 1,592 identities. For face verification, our model was trained with CASIA-Webface

dataset (Yi et al., 2014), which contains 10,575 different identities with 494,414 facial images. Data augmentation with both horizontal flip and random cropping is applied.

2.3 Loss Function

2.3.1 Overall Loss Function

To make outstanding networks that can extract discriminative representation, both inter-class variance and intra-class variance should be considered. Network with large inter-class variance focuses on distinguishable features that are required to verify whether input images are same or not. On the other hand, small intra-class variance implies that latent features of given images are stable regardless of its transformation such as rotation, blurring, etc. In other words, it guarantees the robustness of extracted features.

Our approach aims to minimize intra-class variance and maximize inter-class variance. We implemented both identification and verification to modulate inter-class variance and intra-class variance. The total loss function for our network its loss is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{id}^{h} + \lambda \mathcal{L}_{ve} \tag{2}$$

where $\mathcal{L}_{id}^{h}$, $\mathcal{L}_{ve}$, and λ are identification loss with loss-weighted gradient, verification loss, and the hyper-parameter that modulates ratio between identification loss and verification loss. Identification is performed with softmax classification. Using extracted latent features α over softmax classifier, it is formulated as:

$$\mathcal{L}_{id} = -\sum_{i=1}^{N} [\hat{o}_i \log o_i + (1-\hat{o}_i)\log(1-o_i)] \tag{3}$$

where o_i is the output of softmax classifier, and $\hat{o}_i$ is the desired output value. For identification loss of all training data:

$$\mathcal{L}_{id}^{h} = \frac{1}{M}\sum_{j=1}^{M} L_{id,j}^{h} \tag{4}$$

where M is the number of training examples, and h is hyper-parameter that is used for loss-weighted gradient. This identification task takes a role of making the margins among groups of feature that have the same identity. This role could be interpreted as optimization of inter-class variance.

To achieve smaller intra-class variance, we employed the verification task using $L2$ normalized latent features. In the verification task, we employ the triplet loss as follows:

$$\mathcal{L}_{ve} = \sum_{i=1}^{M} [d(L(a_i), L(p_i)) - d(L(a_i), L(n_i)) + \alpha] \tag{5}$$

where a_i, p_i and n_i are i-th latent feature of anchor, positive, and negative sample respectively. Verification task considers not only samples of target identity but also that of different identities.

2.3.2 Details of Loss-weighted Gradient

Calculating the gradient of a single batch from equation (6), it is formulated as:

$$g(L_{id,j}) = \frac{\partial L_{id,j}}{\partial w} \tag{6}$$

where j implies index of given batch. For weight gradient of the element in single batch with respect to its loss by $L_{id,j}$ raised to the power of h :

$$\begin{aligned} g((L_{id,j})^h) = \frac{\partial (L_{id,j})^h}{\partial w} &= \frac{\partial (L_{id,j})^h}{\partial L_{id,j}} \frac{\partial L_{id,j}}{\partial w} \\ &= h((L_{id,j})^{h-1}) g(L_{id,j}) \end{aligned} \tag{7}$$

Meaning that weight update of individual loss is multiplied by its loss. A batch with a larger loss will be weighted more than one with a smaller loss. To prevent the gradient from getting too small, or having too large value, we implemented $L_{id,j}^{h}$ that has the following gradient:

$$g(L_{id,j}^{h}) = \frac{Mh \cdot L_{id,j}^{h-1}}{\sum_{k=1}^{M} L_{id,k}^{h-1}} \, g(L_{id,j}) \tag{8}$$

Where M is the size of mini batch. Average of $(L_{id,j})^{h-1}$ is divided after calculation of $g(L_{id,j})$. Applying loss weighted gradient on identification, hard-sample for each identity will have majority over updating weights. To show its effect over the network, we conducted additional experiment on MNIST Dataset.

3 EXPERIMENTAL RESULTS

3.1 Loss-weight Gradient on MNIST Dataset

We conducted an experiment with classification network with loss-weight gradient using MNIST dataset (LeCun et al., 1998) to verify the efficiency of the loss-weighted gradient. In this experiment, we visualize activation results of last hidden layer. The dimensionality of last hidden layer is equal to 2, therefore value of the output can be interpreted as coordinates

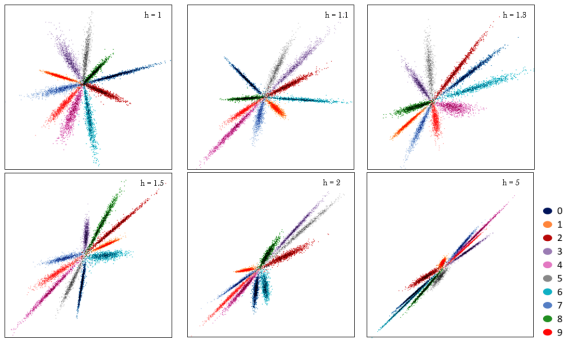

Figure 2: Visualization results on MNIST dataset with respect to h.

in the x-y plane. 10-way softmax classification is performed with cross-entropy loss.

Figure 2 illustrates the visualization result of 2-D latent features on the last hidden layer. By analyzing visualization results of softmax with respect to hyper parameters h, we show that different value of hyper-parameter h result in the different distribution of samples' embedding. The case with $h = 1$, which is general softmax classifier, shows well-spreading shape of embeddings. However, spreads of individual class shrinks when value of h becomes large. Especially in case of $h = 5$, the inter-class distance becomes too small that it is barely possible to perform classification among them. Since unconstrained conditions does not ensure the non-existence of unknown class between known ones, larger h will result in lower accuracy. In our case, we conduct experiments to find appropriate h value to get the best results.

3.2 Face Verification on LFW and YTF Datasets

For model evaluation, Labeled Faces in the Wild (LFW) dataset (Huang et al., 2007) and Youtube Faces (YTF) (Wolf et al., 2011) dataset, known to be exclusive to CASIA-Webface, are used. We followed *unrestricted with labeled outside data* protocol for evaluation. The face verification performance of the proposed model is evaluated on 6,000 of face pairs from LFW dataset, and 5,000 of video pairs from YTF dataset.

We evaluated the proposed model on LFW and YTF dataset in *unconstrained protocol*. Table 1 shows proposed method in comparison with others for LFW and YTF dataset. Each input facial image is resized into 160x160 resolution. Value of hyper-parameter h is assigned to 1.1 which showed the best performance. Although proposed method did not achieve the highest accuracy, it showed comparable performance to models trained with the public dataset. It showed 98.83% of accuracy over LFW dataset, and 92.40% of accuracy over YTF dataset.

Table 1: Accuracy(%) on LFW and YTF dataset. Private dataset type means that used dataset for training is not publicly accessible.

Method	Dataset	Dataset type	LFW	YTF
DeepFace (Taigman et al., 2014)	4M	Private	97.35	91.4
GaussianFace(Lu and Tang, 2015)	850K	Private	98.52	N/A
Facenet (Schroff et al., 2015)	200M	Private	**99.65**	**95.1**
DeepID2+ (Sun et al., 2015)	200K	Private	99.47	93.2
Baidu (Liu et al., 2015)	1.3M	Private	99.13	N/A
Associate-Predict(Yin et al., 2011)	Multi-PIE	Public	90.57	N/A
DDML(combined)(Hu et al., 2014)	LFW-train	Public	90.68	82.34
ConvNet-RBM(Sun et al., 2013)	CelebFaces	Public	92.52	N/A
high-dim LBP(Chen et al., 2013)	WDRef	Public	95.17	N/A
TL Joint Bayesian(Cao et al., 2013)	WDRef	Public	96.33	N/A
FR+FCN(Zhu et al., 2014)	CelebFaces	Public	96.45	N/A
DeepID(Sun et al., 2014)	CelebFaces	Public	97.45	N/A
Liu etal (Liu et al., 2016)	CASIA	Public	**99.10**	**94.0**
Ours(h=1.1)	CASIA	Public	98.83	92.40

3.3 Person Re-identification on ViPeR Dataset and CUHK-01 Dataset

To demonstrate an efficiency of the proposed learning approach in person re-identification problem, we initially conduct the experiment using ViPeR dataset (Gray et al., 2007). The VIPeR dataset contains 632 pedestrian paired images taken from various viewpoints with diverse illumination conditions. The dataset is collected in an academic setting over the course of several months. We also carry out the experiment using CUHK-01 dataset (Li et al., 2012). The CUHK-01 dataset is composed of 971 identities with 4 images for each. Each image is resized to 128x48 resolution. In addition, we conducted additional experiments to get appropriate value of h for loss-weighted gradient. Table 3 shows the accuracies of the proposed model with respect to values of h on CUHK-01 dataset.

We use the one-trial Cumulative Matching Characteristic (CMC) result to compare the proposed method and others. Table 2 shows the comparison results using the ViPeR dataset and CUHK-01 dataset respectively. Due to the lack of experiments about the unconstrained person re-identification, we have compared the proposed method with approaches which focus in constrained condition. Despite the disadvantage of the unconstrained condition, the proposed method has achieved the state-of-the-art performance. The proposed method shows 59.40% and 73.17% matching rates in 1-rank and 5-rank respectively in ViPeR dataset. In CUHK-01 dataset, proposed method achieved 61.65% of 1-ranked matching rate, and the work presented the state-of-the-art performance among the 1-ranked matching rate across the person re-identification methods.

The experimental results show that proposed method outperforms the existing state-of-the-art approaches under the disadvantage of the unconstrained con-

Table 2: Top-ranked matching rates (%) on ViPeR and CUHK-01 dataset. *p*-scale denotes the scale of the set of probe images. The **bolded figures** represent the best values across their rank. Protocol denotes the environmental condition of the experiment.

	ViPeR Dataset						CUHK-01 Dataset					
Method	Protocol	*p*-scale	$r=1$	$r=5$	$r=10$	$r=20$	Protocol	*p*-scale	$r=1$	$r=5$	$r=10$	$r=20$
ℓ_1-norm (Zhao et al., 2013a)	-	-	-	-	-	-	Closed-set	486	10.33	20.64	26.34	33.52
ℓ_2-norm (Zhao et al., 2013a)	-	-	-	-	-	-	Closed-set	486	9.84	19.84	26.42	33.13
SDALF (Farenzena et al., 2010)	Closed-set	316	19.87	38.89	49.37	65.73	Closed-set	486	9.90	22.57	30.33	41.03
CPS (Cheng et al., 2011)	Closed-set	316	21.84	44.00	57.21	71.00	-	-	-	-	-	-
eSDC (Zhao et al., 2013b)	Closed-set	316	26.31	46.61	58.86	72.77	Closed-set	486	19.67	32.72	40.29	50.58
SalMatch (Zhao et al., 2013a)	Closed-set	316	30.16	52.31	65.54	79.15	Closed-set	486	28.45	45.85	55.67	67.95
MLF (Zhao et al., 2014)	Closed-set	316	29.11	52.34	65.95	79.87	Closed-set	486	34.30	55.06	64.96	74.94
LADF (Li et al., 2013)	Closed-set	316	29.34	61.04	75.98	88.10	-	-	-	-	-	-
MFA (Xiong et al., 2014)	Closed-set	316	32.24	65.99	79.66	90.64	-	-	-	-	-	-
kLFDA (Xiong et al., 2014)	Closed-set	316	32.33	65.78	79.72	**90.95**	Closed-set	486	32.76	59.01	69.63	79.18
DeepRank (Chen et al., 2016)	Closed-set	316	38.37	69.22	**81.33**	90.43	Closed-set	486	50.41	**75.93**	**84.07**	**91.32**
Ours(h=1.1)	Open-set	632	**59.40**	**73.17**	78.90	85.55	Open-set	971	**61.65**	74.95	80.62	85.57

Table 3: Accuracies of the proposed model depending on h over CUHK-01 dataset.

h value	$r=1$	$r=5$	$r=10$	$r=20$
1.0	60.06	72.89	78.66	84.02
1.1	**61.65**	74.95	**80.62**	**85.57**
1.2	61.03	74.97	80.00	84.95
1.3	61.03	**75.05**	80.21	84.95
1.4	61.03	74.02	80.21	84.95
1.5	60.72	74.43	79.48	84.23

dition. The interesting point is that other methods tend to have dependency over the dataset that its accuracy changes when test dataset changes. Our method seems to be independent of its test dataset. It implies that our network is well trained to extract more robust and general features that can be found in any different dataset.

4 CONCLUSION

This paper proposes a densely-connected convolution network with loss-weighted gradient. It has shown great performance over both face verification and person re-identification tasks. The presented approach was evaluated by comparing face verification and person re-identification task with other methods. Although it did not show the best performance in face verification among models trained with the public dataset, it showed comparatively good performance with 98.83% of accuracy over LFW dataset. In the case of person re-identification, it outperformed previous state-of-the-art approaches in both CUHK-01 dataset and ViPeR dataset. Moreover, the presented approach showed its superior stability over dataset compared to other methods that the accuracy of other approaches tends to change over datasets.

ACKNOWLEDGEMENTS

This work was supported by the ICT R&D program of MSIP/IITP. [2014-0-00077, Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis]

REFERENCES

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature verification using a" siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, pages 737–744.

Cao, X., Wipf, D., Wen, F., Duan, G., and Sun, J. (2013). A practical transfer learning algorithm for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3208–3215. IEEE.

Chen, D., Cao, X., Wen, F., and Sun, J. (2013). Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3025–3032. IEEE.

Chen, S.-Z., Guo, C.-C., and Lai, J.-H. (2016). Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367.

Cheng, D. S., Cristani, M., Stoppa, M., Bazzani, L., and Murino, V. (2011). Custom pictorial structures for re-identification. In *Bmvc*, volume 2, page 6.

Farenzena, M., Bazzani, L., Perina, A., Murino, V., and Cristani, M. (2010). Person re-identification by symmetry-driven accumulation of local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2360–2367. IEEE.

Gray, D., Brennan, S., and Tao, H. (2007). Evaluating appearance models for recognition, reacquisition, and tracking. In *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, volume 3, pages 1–7. Citeseer.

Hu, J., Lu, J., and Tan, Y.-P. (2014). Discriminative deep metric learning for face verification in the wild. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1875–1882.

Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2016). Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, W. and Wang, X. (2013). Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3594–3601.

Li, W., Zhao, R., and Wang, X. (2012). Human reidentification with transferred metric learning. In *ACCV*.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014). Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 152–159.

Li, Z., Chang, S., Liang, F., Huang, T. S., Cao, L., and Smith, J. R. (2013). Learning locally-adaptive decision functions for person verification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3610–3617. IEEE.

Liu, J., Deng, Y., Bai, T., Wei, Z., and Huang, C. (2015). Targeting ultimate accuracy: Face recognition via deep embedding. *arXiv preprint arXiv:1506.07310*.

Liu, W., Wen, Y., Yu, Z., and Yang, M. (2016). Large-margin softmax loss for convolutional neural networks. In *ICML*, pages 507–516.

Lu, C. and Tang, X. (2015). Surpassing human-level face verification performance on lfw with gaussianface. In *AAAI*, pages 3811–3819.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823.

Sun, Y., Wang, X., and Tang, X. (2013). Hybrid deep learning for face verification. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1489–1496. IEEE.

Sun, Y., Wang, X., and Tang, X. (2014). Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898.

Sun, Y., Wang, X., and Tang, X. (2015). Deeply learned face representations are sparse, selective, and robust. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2892–2900.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al. (2015). Going deeper with convolutions. Cvpr.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708.

Wang, T., Gong, S., Zhu, X., and Wang, S. (2014). Person re-identification by video ranking. In *European Conference on Computer Vision*, pages 688–703. Springer.

Wolf, L., Hassner, T., and Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE.

Xiong, F., Gou, M., Camps, O., and Sznaier, M. (2014). Person re-identification using kernel-based metric learning methods. In *European conference on computer vision*, pages 1–16. Springer.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*.

Yin, Q., Tang, X., and Sun, J. (2011). An associate-predict model for face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 497–504. IEEE.

Zhao, R., Ouyang, W., and Wang, X. (2013a). Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535.

Zhao, R., Ouyang, W., and Wang, X. (2013b). Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593.

Zhao, R., Ouyang, W., and Wang, X. (2014). Learning mid-level filters for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 144–151.

Zhu, Z., Luo, P., Wang, X., and Tang, X. (2014). Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*.

APPENDIX

	Convolutional Layer			DenseBlock		Inception				
Layers	Input	Kernel	Stride	Layer	Growth	L1	L2	L3	L4	num(Parameters)
Conv1	160x160x3	3x3	2							0.9K
Conv2	79x79x32	3x3	1							9.2K
Conv3	77x77x32	3x3	1							18.4K
MaxPool	77x77x64	3x3	1							0
Conv4	75x75x64	1x1	1							5.1K
Conv5	75x75x80	3x3	1							138.2K
Conv6	73x73x192	3x3	2							442.4K
DenseBlock	36x36x256			6	12					148.4k
Inception_a	36x36x164					192	256	256	384	1.4M
DenseBlock	17x17x804			8	24					1.3M
Inception_b	17x17x498					256	384	256	256	3.0M
DenseBlock	8x8x1394			6	12					2.6M
DenseBlock	8x8x1466			6	12					2.8M
AvgPool										0
FC	1x1x1538									196.9K
Concat										0
Softmax	128									1.4M
Total										13.5M

Parameter details of proposed network. Size of the input image is 160x160x3. Value of *Layer* and *Growth* in *DenseBlock* implies the number of sequential convolution layer and depth of output layer respectively. Value of parameters in *Inception_a* and *Inception_b* implies depth of filters in the first layer, and the second layer in inception block respectively.