Discovering Identity Specific Activation Patterns in Deep Descriptors for Template Based Face Recognition

Claudio Ferrari, Stefano Berretti, Alberto Del Bimbo
Media Integration and Communication Center, University of Florence, Italy

***Abstract*— The majority of recent face recognition systems are based on Deep Convolutional Neural Networks (DCNNs). These networks are trained on massive amounts of face images so as to learn a compact representation *(deep descriptor)* aimed at capturing the identity information. Recognition is then performed by computing some similarity (or distance) measure between descriptors. However, in practice, descriptors encode also other *intra-class* variabilities such as pose and expressions. This well-known problem is usually addressed by designing specific loss-functions or metric learning modules such that the learned descriptors maximize the *inter-class* (identity) distances and minimize the *intra-class* differences in the feature space. We tackle this problem from a different perspective by observing that descriptors associated with images of the same subject, on average, share similar patterns in the highest activation units. We demonstrate this assumption by showing that improved accuracy can be obtained in a template-based recognition scenario by retaining the descriptor bins with the average highest activation, and dropping all the others to zero. These activation patterns are also employed to build identity-representative binary masks that are effectively used in place of the descriptors to match templates. We investigate this strategy by performing experiments on the IJB-A dataset, and show that it can significantly boost the recognition accuracy.**

I. INTRODUCTION

Among different biometric techniques, face recognition has the desirable characteristic of being non-intrusive and to not explicitly require user cooperation. These are mostly the main reasons for the success of face recognition as also evidenced by the increasing demand for surveillance systems that can operate in real contexts. For many years, the majority of face recognition methods made use of hand-crafted features (also referred to as "shallow" features) to capture patterns and recurrent structures from the image pixels. However, the advent of Deep Convolutional Neural Network (DCNN) architectures has radically changed the scenario in face recognition [1]. One substantial innovation of DCNNs is the idea of letting the deep architecture to automatically discover low-level and high-level representations from labeled and/or unlabeled training data, which can then be used for detecting and/or classifying the underlying patterns. CNNs have found effective application and breakthrough results have been obtained using such technology on most of the existing face recognition benchmark datasets [2], [3], [4], [5]. To further push the challenges, new datasets and protocols have then been collected and proposed; the trend has moved towards more realistic scenarios in which recognition has to be performed between *templates*, *i.e.*, sets of images or full video sequences or both [6], [7], each collected with the specific intent of including a considerable number of different individuals in a large variety of conditions. While, on the one hand, the availability of more than one image per subject brings richer information, on the other, it makes the recognition harder because of the increased number of requested comparisons. A major challenge in order to perform accurate recognition in such conditions is to learn highly discriminative face descriptors that effectively capture the identity information regardless of other factors such as pose or expression variations. Such nuisances, along with image-level external factors like low resolution or blurring, indeed can impair the recognition in many circumstances. Most of the effort to solve the above problems has been put in designing specific loss functions or metric learning solutions so as to guide the networks in learning a discriminative embedding for which descriptors of the same subject lie close to each other and far apart from other persons' descriptors in the feature space.

978-1-7281-0089-0/19/$31.00 ©2019 IEEE

A. Motivation and Our Contribution

In this paper, we tackle the problem from a different perspective and propose a preliminary investigation aimed at recovering the identity-related information directly from the deep descriptor. We build our assumption considering that face recognition networks are trained as identity classifiers under the supervision of the cross entropy loss, which tries to maximize the conditional probability of the training samples over the real distribution. In a previous work, Rajan *et al.* [8] already investigated on the effect that this supervisory loss function has on the deep descriptors; they showed that the $L2$-norm of the resulting descriptors is informative of the quality of the face. Indeed such loss tends to generate descriptors with high $L2$-norm for the "easy" instances, *e.g.*, good quality frontal faces, while "hard" examples, *e.g.*, blurry or low resolution images result in a vector with low $L2$-norm, as depicted in Fig. 1. Intuitively, this means that classification uncertainty at training time is modeled by low-valued activations. Based on this, to compensate this behavior and boost recognition accuracy, they added an $L2$-constraint to the face descriptors, which restricts them to lie on a hypersphere of fixed radius α. As described in [8], α has to be large (in the order of 50) both because the hypersphere needs to have sufficiently large surface for embedding the features, and because a larger norm implies a higher probability of correct identity classification. The latter statement is supported by the evidence that good descriptors have higher norm and tend to have evident high-valued spikes

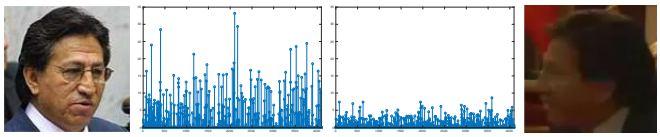

Fig. 1. "Good" (left) and "bad" (right) examples. The descriptor associated to the good example has way larger average activation values and evident spikes, while the other shows the opposite. The two plots have same scale.

in correspondence of some bins, with a large magnitude gap between those bins and the others (Fig. 1, left). Thus, our hypothesis is that these spikes might encode the identity information we need. Further, we focused our attention on the fact that the cosine similarity between descriptors, which is widely recognized as a very effective similarity measure for high dimensional vectors, is a measurement of orientation rather than magnitude; this implies that two vectors with different magnitude pointing in the same direction, will still have high similarity. Thus, under this point of view, we can state that the activations pattern within the descriptor is of fundamental importance for a correct and precise matching. Putting all together, in this work, we investigate if such identity-specific recurrent patterns can be found considering the descriptor bins with highest activation.

The rest of the paper is organized as follows: In Section II, we summarize the works in the literature that are closer to the proposed method; In Section III, we present the method used for selecting the most relevant features from the descriptors and how it can be injected into a recognition pipeline; Results obtained in a comprehensive set of experiments are reported in Section IV; Discussion, limitations and conclusions are sketched in Section V.

II. RELATED WORK

Works that are related to our approach can be categorized according to the fact they *(i)* define specific loss-functions or metric learning solutions, or *(ii)* use template pooling or image attention from image sets.

In face verification, a typical pipeline includes training a deep network for subject classification with softmax-loss, using the penultimate layer output as the feature descriptor, and generating a cosine similarity score given a pair of face images. However, the softmax-loss function does not optimize the features to have higher similarity score for positive pairs and lower similarity score for negative pairs, which can lead to a performance gap. One possible solution to solve this problem is that to use pairs of face images as input to the training algorithm to learn a feature embedding, where positive pairs are closer and negative pairs are far apart. Following this idea, solutions have been proposed that use siamese networks with contrastive-loss [9], discriminative deep metric with a margin between positive and negative face pairs [10], triplet-loss metric learning [11], and center-loss [12]. Other solutions used the face images along with their subject labels and train a DCNN with softmax-loss to learn discriminative identification features in a classification framework [2], [4], [13]. These features are later used either to directly compute the similarity score for a pair of faces or to train a discriminative metric embedding [14], [15]. Training the network for joint identification-verification task is another possible strategy [12], [13], [16]. Recently, a few algorithms have used feature normalization during training to improve performance. Liu *et al.* [17] proposed the angular softmax (A-Softmax) loss that enables CNNs to learn angularly discriminative features. Geometrically, the A-Softmax loss can be viewed as imposing discriminative constraints on a hyper-sphere manifold, which intrinsically matches the prior that faces also lie on a manifold. Hasnat *et al.* [18] used a special case of batch normalization technique to normalize the feature descriptor before applying the softmax-loss.

Other methods try to perform a feature aggregation based on template pooling and attention. Hassner *et al.* [19] proposed a method to both increase the recognition accuracy and reduce the computational cost of template matching. To do this, they leveraged on average pooling of face photos. They showed how the space of a templates images can be partitioned and then pooled based on image quality and head pose and the effect this has on accuracy and template size. Yang *et al.* [20] presented a Neural Aggregation Network (NAN) for video face recognition. The network takes a face video or face image set of a person with a variable number of face images as input, and produces a compact, fixed-dimension feature representation. The whole network is composed of two modules: the feature embedding module and the aggregation module. Tran *et al.* [21] proposed a Disentangled Representation learning-Generative Adversarial Network (DR-GAN) that generates a weight for each input using a (sigmoid) gating function, and aggregates from the multiple inputs using a weighted average. The encoder-decoder structure of the generator allows DR-GAN to learn a generative and discriminative representation, in addition to image synthesis. This representation is explicitly disentangled from other face variations such as pose, through the pose code provided to the decoder and pose estimation in the discriminator. In this way, DR-GAN can take one or multiple images as the input, and generate one unified representation along with an arbitrary number of synthetic images. Ding *et al.* [22] proposed to build a large-scale face recognizer, which is capable to fight off the data imbalance difficulty. To seek a more effective general classifier, they developed a generative model to synthesize meaningful data for one-shot classes by adapting the data variances from other normal classes. They formulated conditional GANs and the general Softmax classifier into a unified framework. Such a two-player minimax optimization can guide the generation of more effective data, which benefit the classifier learning for one-shot classes. In [23], Xie and Zisserman also focused on set-based face recognition using an attention-based mechanism, and adding ideas from relation/metric learning [24], [25]. In contrast to conventional solutions, where the set-wise feature descriptor is computed as an average of the descriptors from individual face images within the set, a neural network architecture is proposed that learns to aggregate based on both "visual" quality (resolution,

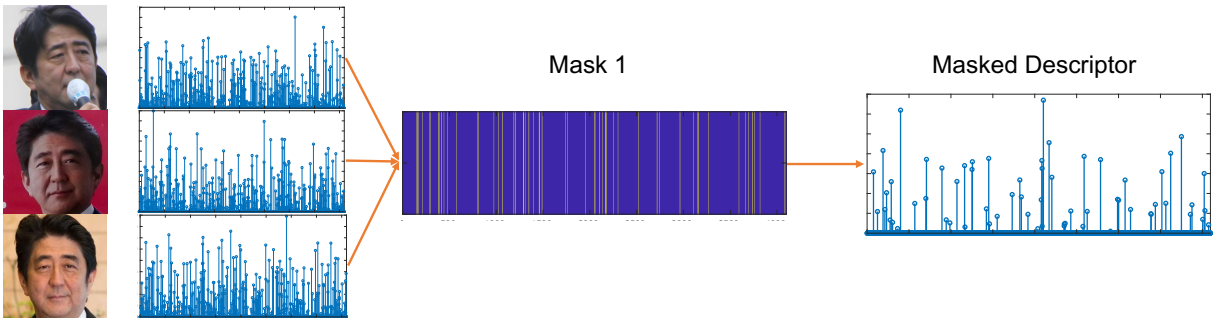

Fig. 2. Overview of our approach. The template descriptors are used to build a binary mask M, which is then employed to filter-out low activations.

illumination), and "content" quality (relative importance for discriminative classification). To this end, a Multi column Network (MN) is proposed that takes a set of images (the number in the set can vary) as input, and learns to compute a fix-sized feature descriptor for the entire set.

III. ACTIVATION PATTERN ESTIMATION

Grounding on the assumptions and observations expounded so far, we developed an intuitive and straightforward approach to estimate identity-specific activation patterns from face descriptors in the case a template of images or video frames is available for each subject. Given a template $\mathbf{T}$ of n descriptors $\mathbf{d} \in \mathbb{R}^k$ of a single individual, our strategy, depicted in Fig. 2, consists of four steps:

1) Compute the average, un-normalized, descriptor $\overline{\mathbf{d}}$ from the template;
2) Find the m bins of $\overline{\mathbf{d}}$ with highest activation;
3) Build a binary mask $\mathbf{M} \in \{0,1\}^k$ of the same dimension of $\mathbf{d}$, in which the elements corresponding to the m most active bins of $\mathbf{d}$ have value 1;
4) Mask out the m bins from the original descriptors as

$$\hat{\mathbf{d}}_i = \mathbf{d}_i \wedge \mathbf{M} \,. \tag{1}$$

Each new descriptor $\hat{\mathbf{d}}_i$ of the template has the same dimension of the original one $\mathbf{d}_i$, but only m bins are $\neq 0$, while all the other $k-m$ bins are dropped to zero. Note that considering the un-normalized descriptors to estimate $\overline{\mathbf{d}}$ is important in as much as, in doing so, we intrinsically weigh down the contribution of ambiguous descriptors, *i.e.*, with low $L2$-norm, performing a sort of automatic image selection. The procedure is applied separately to each template.

The template-based recognition is finally performed following a standard pipeline and the guidelines of [26]; in particular, we chose to employ the *min-mean* cosine distance criterion, as it showed to be a simple yet robust measure for matching image sets. The latter sums up the minimum and the average distance between templates' descriptors. Note that, differently from [26], we do not apply PCA.

Within the recognition pipeline, our proposed method is evaluated in two different ways: *(i)* the masked descriptors are used in place of the original ones; *(ii)* the binary masks are used as identity representatives and used in place of the descriptors to match the templates. The distance between two templates $(\mathbf{T}_i, \mathbf{T}_j)$, then becomes equal to the distance between the two masks $(\mathbf{M}_i, \mathbf{M}_j)$ associated to them.

IV. EXPERIMENTAL EVALUATION

We evaluated our proposed approach on the IARPA Janus Benchmark-A (IJB-A) [6] dataset, using the VggFace model [4] for extracting 4096-dimensional face descriptors.

The IJB-A dataset includes face imagery coming both as still images and video frames, captured under severe variations of imaging conditions, focusing on the extreme cases. The dataset comprises $25,800$ images of 500 individuals. Two main protocols are defined: face identification (1:N) and face verification (1:1). In both, the identities to be matched or retrieved are expressed by means of *templates*, *i.e.*, collections of images (or video frames) of the same individual. Results are reported following the dataset convention; for the identification protocol, we additionally report TAR@FAR = $10^{-4}/10^{-5}$, which is similar to the True Positive at a given False Positive Identification Rate (TPIR@FPIR), but in a closed-set scenario.

In the first experiment, we compared the accuracy of the original descriptors against the masked ones. Here, the number of bins to be retained m is a hyper-parameter that must be defined a priori; in Table I, we report results for different values of m, ranging from 50 up to 1024, value after which accuracies begin to degrade. Results show that by selecting a small subset of the activations, we considerably boost the performance; best results are obtained with a very small amount of active bins (150, 250 out of 4096). To further verify our assumptions, in the last row of Table I, we report results obtained by selecting a random subset of $m = 1024$ descriptor bins. As expected, in this case results drop dramatically, providing great support to our hypothesis for which the identity information is encoded in just few specific activations. We do not report random selections for different m values as the outcome does not significantly vary. Different behaviors are observed for the identification and verification protocol; indeed, a large increase of accuracy is obtained for the identification protocol, while performance in verification increase slightly. A deepened analysis disclosed that the reason is the different template sizes. Indeed, in the first case, the average number of gallery/probe images per template is, respectively, around 30 and 120, while for the verification protocol is around 30 and 7. Reasonably,

TABLE I

RESULTS ON IJB-A USING MASKED DESCRIPTORS FOR DIFFERENT m VALUES. BEST RESULTS IN BOLD

	Identification 1:N						Verification 1:1		
m	FPIR=10^{-1}	FPIR=10^{-2}	FAR=10^{-4}	FAR=10^{-5}	Rank@1	Rank@10	FAR=10^{-1}	FAR=10^{-2}	FAR=10^{-3}
Original ($m = 4096$)	0.641	0.352	0.379	0.197	0.885	0.976	0.953	0.797	0.539
50	0.724	0.530	0.517	0.347	0.906	0.984	0.928	0.740	0.495
150	0.768	0.568	**0.585**	**0.407**	0.952	0.992	0.944	0.802	**0.579**
250	**0.791**	**0.583**	0.573	0.382	**0.954**	**0.996**	0.951	0.812	0.572
500	0.739	0.497	0.494	0.302	0.940	0.992	0.953	**0.816**	0.574
1024	0.674	0.391	0.416	0.221	0.904	0.985	**0.954**	0.802	0.550
Random	0.172	0.023	0.062	0.018	0.509	0.870	0.449	0.133	0.030

TABLE II

RESULTS ON IJB-A USING THE BINARY MASKS IN PLACE OF THE DESCRIPTORS FOR DIFFERENT m VALUES. BEST RESULTS IN BOLD

	Identification 1:N						Verification 1:1		
m	FPIR=10^{-1}	FPIR=10^{-2}	FAR=10^{-4}	FAR=10^{-5}	Rank@1	Rank@10	FAR=10^{-1}	FAR=10^{-2}	FAR=10^{-3}
Original Descriptors	0.641	0.352	0.379	0.197	0.885	0.976	0.953	0.797	0.539
150	0.901	0.797	0.772	0.676	0.980	0.995	0.934	0.777	0.578
250	0.917	0.785	0.76	**0.731**	0.981	0.996	0.942	0.799	0.564
500	0.904	0.783	**0.789**	0.678	**0.986**	**0.999**	0.953	0.803	0.540
Template Adaptation [27]	0.882	0.774	–	–	0.928	0.986	0.979	0.939	0.836
All-In-One+TPE [28]	0.887	0.792	–	–	0.947	0.988	0.976	0.922	0.823
NAN [20]	0.917	0.817	–	–	0.958	0.986	0.978	0.941	0.881
$L2$-S+TPE [8]	**0.956**	**0.915**	–	–	0.973	0.988	**0.984**	**0.970**	**0.943**

templates with many images increase the chance of including good examples and the statistical robustness of the selection as well. Another aspect that considerably affects the effectiveness of our strategy is the variability within the face images; we found that if a template is not sufficiently variegated, *e.g.*, a short video sequence, the selection loses its potential to some extent. Even though this might represent a limitation of this approach, performance still gain a slight improvement. In any case, interesting and innovative solutions to overcome this limit might be found; for example, templates could be augmented with novel images exploiting a generative model as the one in [29], which showed great potential for face recognition applications.

In the second experiment, the binary masks are used in place of the descriptors to match the templates. Note that, in this case, we associate a single binary mask to each template, which significantly simplifies and speeds up the matching. This experiment was conducted because, in the previous approach, ambiguous descriptors could still impair the recognition. Intuitively, even if we filter out non-discriminative activations, if the original descriptor was unable to capture the identity traits, the filtered descriptor will hardly be effective as well. Table II shows that the binary masks are actually extremely effective as representatives of the identities. Similarly to the previous experiment, the effect is way more evident for the identification protocol, where we obtain a remarkable 98.6 accuracy at rank-1 using only 250 activations and a huge improvement in TAR at very low FAR and when false positives, *i.e.*, out of gallery subjects are involved. The large improvement obtained for very low FAR values makes sense since the binary masks exclude non-informative dimensions and prevent mismatches resulting from noisy descriptors. Table II includes also results from recent state-of-the-art approaches in order to assess the competitiveness of our solution; from such comparison, we can observe that it provides results in line with the state-of-the-art. However, note that the compared methods either perform a Triplet Probabilistic Embedding (TPE) of the descriptors [8], [28] or fine-tune the model [20] on the training splits of the IJB-A, while we simply use the pre-trained network as is.

V. CONCLUSIONS

In this paper, we have proposed a preliminary investigation and a intuitive and straightforward approach aimed at discovering identity-specific patterns in the activations of face descriptors. We showed that the identity related information is encoded in very few bins of the descriptors, which correspond to the ones with highest activation. With this approach, we can significantly boost the performance of a standard network architecture. We also showed that the binary masks derived from the bins selection can be used as identity-representatives for effectively matching templates. On the other hand, we found that an evident limitation is that the number of available descriptors per identity can somewhat impair the selection and limit the accuracy gain. Despite this, the outcomes reported in this paper are valuable in as much as they shed some light on the learned face representation, and can be useful to find new architectural and learning solutions. As future work, we aim to extend our investigation to different datasets, also finding a solution to modify and integrate the proposed module directly into the network architecture.

VI. ACKNOWLEDGMENTS

The Titan Xp used for this research was donated by the NVIDIA Corporation.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Int. Conf. on Advances in Neural Information Processing Systems (NIPS)*, 2012.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[3] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[4] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conf. (BMVC)*, 2015.

[5] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[6] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, M. Burge, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[7] C. Ferrari, S. Berretti, and A. Del Bimbo, "Extended youtube faces: a dataset for heterogeneous open-set face identification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3408–3413.

[8] R. Ranjan, C. D. Castillo, and R. Chellappa, "L_2-constrained softmax loss for discriminative face verification," *CoRR*, vol. abs/1703.09507, 2017. [Online]. Available: http://arxiv.org/abs/1703.09507

[9] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, June 2005, pp. 539–546.

[10] J. Hu, J. Lu, and Y. Tan, "Discriminative deep metric learning for face verification in the wild," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2014, pp. 1875–1882.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 815–823.

[12] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conf. on Computer Vision (ECCV)*, vol. 9911, 2016, pp. 499–515.

[13] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 2892–2900.

[14] J.-C. Chen, V. M. Patel, and R. Chellappa, "Unconstrained face verification using deep CNN features," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2016.

[15] S. Sankaranarayanan, A. Alavi, C. D. Castillo, and R. Chellappa, "Triplet probabilistic embedding for face verification and clustering," in *IEEE Int. Conf. on Biometrics Theory, Applications and Systems (BTAS)*, 2016.

[16] Y. Wen, Z. Li, and Y. Qiao, "Latent factor guided convolutional neural networks for age-invariant face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 4893–4901.

[17] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6738–6746.

[18] A. Hasnat, J. Bohne, J. Milgram, S. Gentric, and L. Chen, "Deepvisage: Making face recognition simple yet with powerful generalization skills," in *IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, October 2017, pp. 1682–1691.

[19] T. Hassner, I. Masi, J. Kim, J. Choi, S. Harel, P. Natarajan, and G. Medioni, "Pooling faces: Template based face recognition with pooled face images," in *IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016, pp. 127–135.

[20] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural aggregation network for video face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5216–5225.

[21] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning gan for pose-invariant face recognition," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 1283–1292.

[22] Z. Ding, Y. Guo, L. Zhang, and Y. Fu, "One-shot face recognition via generative learning," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*, May 2018, pp. 1–7.

[23] W. Xie and A. Zisserman, "Multicolumn networks for face recognition," in *British Machine Vision Conf. (BMVC)*, Sept. 2018.

[24] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2017, pp. 4967–4976.

[25] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2018, pp. 1199–1208.

[26] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo, "Investigating nuisances in DCNN-based face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5638–5651, 2018.

[27] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, "Template adaptation for face verification and identification," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*, 2017.

[28] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *IEEE Int. Conf. on Automatic Face & Gesture Recognition (FG)*, 2017, pp. 17–24.

[29] J. Zhao, L. Xiong, P. K. Jayashree, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, "Dual-agent gans for photorealistic and identity preserving profile face synthesis," in *Advances in Neural Information Processing Systems*, 2017, pp. 66–76.