RECOGNIZING MINIMAL FACIAL SKETCH BY GENERATING PHOTOREALISTIC FACES WITH THE GUIDANCE OF DESCRIPTIVE ATTRIBUTES

Xiao Yang[1] *Hang Su*[2] *Qin Zhou*[1] *Xinzhe Li*[1] *Shibao Zheng*[1]

[1] Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Tsinghua National Lab for Information Science and Technology, Beijing, China

ABSTRACT

Cross-modal sketch-photo recognition is of vital importance in law enforcement and public security. Most existing methods are dedicated to bridging the gap between the low-level visual features of sketches and photo images, which is limited due to intrinsic differences in pixel values. In this paper, based on the intuition that sketches and photo images are highly correlated in the semantic domain, we propose to jointly utilize the low-level visual features and high-level facial attributes to enhance the representation ability of sketches. More specifically, a Multi-Modal Conditional GAN (MMC-GAN) is proposed to generate face images for further face recognition based on the generated images. During training, an identity-preserving constraint is further introduced to improve the discriminative ability of the synthetic images. Extensive experiments demonstrate that the effectiveness of attribute-aided face synthesis and recognition.

Index Terms— Face synthesis; Face recognition; Sketch; Attributes

1. INTRODUCTION

Recognizing a face from its facial sketch is an important, yet challenging problem in the face recognition community. It has wide practical applications in security and law enforcement, since the photo image of a suspect is unavailable in many cases. A face sketch captures distinctive characteristics of human faces, while the natural language descriptions with high-level semantic meaning (gender, hair color, etc.) further complement the missing texture information for the sketches.

During the past years, significant attempts have been made in sketch-photo matching [1, 2, 3], which aim to bridge cross-modality gap by allowing heterogeneous imagery to be compared for recognition, e.g. using multi-scale Markov Random Fields (MRF) [4] and one corresponding template sketch [5]. Recently, photorealistic face synthesis based on sketch [6] has been proposed to generate face images from sketches based on GAN [7], considering its powerful synthesis capabilities [8, 9, 10]. The authors further proposed to generate images from attributes [11] based on Variational Auto-Encoder (VAE) [12] or GANs [13, 14]. Nevertheless, these methods are ill-posed problem since they fail to integrate the conditional control (hair color, skin color, etc.). How to jointly utilize the heterogeneous data to improve the face recognition performance, is however far from solved.

To address the aforementioned issues, we propose to generate and recognize faces based on a proposed conditional GAN [8] with facial sketches and binary attributes, endowing multi-modal information into a hidden incorporation. More specifically, the global structures and local details of a face are respectively encoded via different channels of the generator, and then fused together with the guidance of the corresponding descriptive attributes. Since there is no existing dataset focusing on multi-modal information, we construct two new facial sketch-attribute datasets named Sketch-CelebA and Sketch-LFWA using face attribute dataset CelebA [15] and LFW [16]. Particularly, we use the minimal facial sketches by plotting 68 facial key landmarks [17] in original pixel color, which is much more simple but effortless to generate than hand-drawn pencil sketch by forensic artists. The synthetic sketch contains partial color information of original face image, which is helpful for generating faces with realistic skin color and race information. Considering that a synthetic sketch lacks texture information, we propose to utilize descriptive attributes to serve as additional conditional complementarity to provide detailed information. Rather than integrate all the attributes (e.g., 40 attributes for CelebA), we manually select 19 representative attributes for each image to complement the sketch. By simply altering the attribute value, the network integrates new condition information to generate a face that is consistent with the modified attributes.

In summary, we propose a novel structure of Multi-Modal Conditional GAN (MMC-GAN) to generate photorealistic facial images from sketch and high-level descriptive information, based on which face recognition can be effectively implemented. To the best of our knowledge, this is the first attempt to perform face image generation and recognition by jointly utilizing multi-modal information, which is of good practical value for real-world scenarios. Extensive experiments demonstrate that the generated faces are much easier to be recognized for users than the original sketches.

This work was supported by the NSFC(No.61671289 and No.61571261) and STCSM Grant 15DZ2270400.

2. METHODOLOGY

In this section, we present the Multi-Modal Conditional GAN (MMC-GAN) to jointly exploit the heterogeneous data (minimal landmark sketches and descriptive binary attributes) for photorealistic face synthesis. The aim of the proposed MMC-GAN is to automatically learn the distribution $p\left(X_i^R | X_i^S, \mathbf{y}_i\right)$ to generate real image X_i^R with the aid of corresponding visual sketch X_i^S and the related binary attributes y_i. Mathematically, the objective function is formulated as follows:

$$\begin{aligned}&\min_G \max_D \mathbb{E}_{X_i \sim P_{data}, \mathbf{y}_i \sim P(\mathbf{y})}[log(D(X_i^R, X_i^S, \mathbf{y}_i))]+\\&\mathbb{E}_{X_i \sim P_{data}, \mathbf{y}_i \sim P(\mathbf{y})}[log(1 - D(G(X_i^S, \mathbf{y}_i), X_i^S, \mathbf{y}_i)]+\\&\mathbb{E}_{X_i \sim P_{data}, \bar{\mathbf{y}}_i \sim P(\bar{\mathbf{y}})}[log(1 - D(X_i^R, X_i^S, \bar{\mathbf{y}}_i)],\end{aligned} \tag{1}$$

where $\bar{\mathbf{y}}_i$ is the false attribute encoding by randomly flipping the attribute encoding $\mathbf{y}_i$. The generator G tries to generate real data X_i^R given sketch data X_i^S along with the corresponding $\mathbf{y}_i$. The discriminator D tries to distinguish between fake triplets $\{(G(X_i^S, \mathbf{y}_i), X_i^S, \mathbf{y}_i), (X_i^R, X_i^S, \bar{\mathbf{y}}_i)\}$ and real triplets $\left(X_i^R, X_i^S, \mathbf{y}_i\right)$.

The overall architecture of MMC-GAN is presented in Fig. 1, which consists of a face generator G and a conditional discriminator D. The detailed structure of each part is presented as follows.

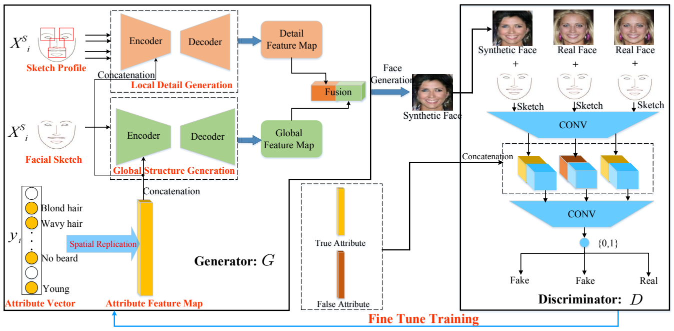

Fig. 1: The architecture of the proposed MMC-GAN for photorealistic face generation, which includes modules of face generator and conditional discriminator. The generator integrates the minimal sketch and descriptive attributes into a hidden incorporation, which complements each other for face synthesis. Note that the global and detail information of a facial sketch is embedded into the feature space via different feature generators, respectively. As for the discriminator, we first produce positive and negative inputs by incorporating the synthetic and real faces with the alternative facial attributes. The generated patches are fed into the conditional discriminator to distinguish the real/fake faces, with the loss training fine the generator.

2.1. Multi-Modal Generator.

We adopt the two-pathway structure proposed in [18] as the baseline of our Multi-Modal Generator, considering its robustness against face rotation during face image generation. Specifically, the generator is decoupled into two parts: G_g and G_l, with G_g focusing on recovering the global face structure, and G_l for reconstructing rich facial details of four important landmarks including left eye, right eye, nose and mouth. Note that we generate the local face patches by putting the corresponding key landmarks (e.g., left eye, right eye, nose and mouth) at the center of the patches. The detail feature map is formed by integrating feature maps of the four landmark generators in a max-value pooling method. The feature maps generated by G_g and G_l are then concatenated together for further face image generation. The combination of the global and local based subnetworks make the generated images more realistic, however, merely utilizing the low-level visual features from sketches will still lead to arbitrary results since different face images can have quite similar sketches.

To address this issue, we invoke additional attribute encoding to complement the synthetic image. More specifically, the attribute information is encoded in a N dimensional binary vector y_i, with each element of y_i representing the presence or absence of a specific attribute. The adopted attributes are presented in Tab. 1. For G_g, the attribute encoding is firstly spatially replicated to form a $L_g \times L_g \times N$ feature map, which has the same height and width as the feature map of the first layer. Then the attribute feature map is concatenated with the feature map of the first layer by channel to generate a facial image after a series of down-sampling and up-sampling blocks. Similarly, partial attributes (e.g., Bags under eyes, Bushy eyebrows) are integrated with G_l. Different generators separately take advantage of corresponding attribute encoding, making generation more reasonable and efficient.

Table 1: Facial attributes grouping.

Facial Part	Facial Attributes
Hair	Black hair, Blond hair, Brown hair, Gray hair, Bald, Wavy hair, Straight hair, Receding hairline, Bangs
Beard	No beard, Goatee, 5 o'clock shadow, Mustache, Sideburns
Wearings	Eyeglasses, Hat
Detail	Bags under eyes, Bushy eyebrows
Appearance	Young, Male

2.2. Conditional Discriminator.

The conditional discriminator aims to distinguish between the real pairs and synthetic pairs, with the loss affecting the generator. The synthetic image X_i^{R*} generated by multi-modal generator and the ground-truth image X_i^R are concatenated with sketch X_i^S by channel to produce positive pairs and negative pairs. The image pairs along with the related attributes are then fed into the network the same way as in the generator. Note that the network may easily ignore the information if we always set the attributes as the true settings. Therefore, we construct an additional group of fake triplets $\left(X_i^R, X_i^S, \bar{\mathbf{y}}_i\right)$, to make sure that the attribute information is functional during the image generation. Here $\bar{\mathbf{y}}_i$ is the false attribute encoding by randomly flipping the binary value of $\mathbf{y}_i$. We use this discriminator to predict different triplets, which is used to fine tune the generator.

2.3. Optimization

In this section, we introduce the adopted loss function for identity-preserving face generation.

Content loss. The pixel-wise loss is calculated as:

$$L_c = \frac{1}{M_c} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{k=1}^{C} |X^R_{i,j,k} - G(X^S, \mathbf{y})_{i,j,k}|, \tag{2}$$

where $M_c = W \times H \times C$. W and H denote dimensions. We adopt the L_1 norm since the L_2 norm lacks the high frequency information, leading to smooth textures in face generation.

Attribute preserving loss. Reconstructing attributes is a very vital part for identity-preserving face generation. High-level attribute representation in the network efficiently reflects a particular attribute configuration. Given face images X_i and corresponding binary attributes $\mathbf{y}_i$, an attribute encoder network (AEnet) is learned beforehand by minimizing the cross entropy loss as follows:

$$L = \sum_{k=1} \mathbf{y}_i^k \log p(\mathbf{y}_i^k = 1|X_i) + (1 - \mathbf{y}_i^k) \log(1 - p(\mathbf{y}_i^k = 1|X_i)), \tag{3}$$

where $p(\mathbf{y}_i^k = 1|X_i)$ is a sigmoid function, implying the probability of the existence of the k-th attribute in the i-th face image X_i.

Attribute preserving loss in the proposed MMC-GAN is then defined as the L1 distance between the features of generated image $G(X^S, \mathbf{y})$ and the real image X^R extracted from the AEnet.

$$L_{ap} = \frac{1}{M_{ap}} \sum_{i=1}^{K_n} \sum_{j=1}^{H_n} \sum_{k=1}^{C_n} |\phi_n(X^R)_{i,j,k} - \phi_n(G(X^S, \mathbf{y}))_{i,j,k}|, \tag{4}$$

where $M_{ap} = W_n \times H_n \times C_n$, and ϕ_n describes the last convolution feature map after activation of the attribute network. W_n and H_n denote dimensions of the feature map.

Adversarial loss. The adversarial loss is also included to maximize the probability that the synthesized image is mistake as a real image by the discriminator,

$$L_{gan} = \frac{1}{N} \sum_{i=1}^{N} -\log D(G(X_i^S, \mathbf{y}_i)), \tag{5}$$

where $D.$ is the probability that $G(X_i^S, \mathbf{y}_i)$ is regarded as a real facial image.

During the training phase, the generator minimizes the objective by summing up the aforementioned loss as

$$L_O = L_{gan} + \lambda_1 L_c + \lambda_2 L_{ap}, \tag{6}$$

where λ_1 and λ_2 are the balanced parameters of the corresponding terms. The proposed MMC-GAN is trained end-to-end with ADAM, except for the attribute encoder, which is pre-trained for attribute classification and fixed afterwards. Empirically, the parameters are set as $N_d = 19$, $L_g = 64$, $L_d = 64$, $\lambda_1 = 10$ and $\lambda_2 = 0.1\lambda_1$.

3. EXPERIMENTS

3.1. Dataset and Evaluation Metrics

Dataset. We construct new datasets Sketch-CelebA and Sketch-LFWA using two public face attribute datasets, namely CelebA [15] and LFW [16]. Different from previous hand-drawn method and stylization algorithm [19], we use the minimal facial sketches by plotting 68 facial key landmarks. The ERT method [17] is adopted to perform landmark detection, which is implemented with Dlib [20] on the 300-W dataset [21]. We manually select 19 representative attributes for each image to complement the sketch. Therefore, Sketch-CelebA and Sketch-LFWA contain over 100,000 subjects and 10,000 subjects, respectively. Each subject has a minimal sketch along with 19 facial key attributes.

Setting and metrics. Sketch-CelebA is divided into two parts, where 50,000 images are used as our training set to optimize the network, and the rest of the images without having overlaps with same person are used to test synthetic performance. Our recognition task is evaluated on Sketch-LFWA by using the standard protocol [16] and test on 10 folders each with 300 same-person pairs and 300 different-person pairs; and each pair consists of a real image and a sketch image along with attributes. Notably, when being evaluated on Sketch-LFWA, all algorithms are only trained on images from original images of Sketch-CelebA.

Comparison methods. We evaluate the alternative face synthesis algorithms including U-Net [22] and pix2pix [8] for comparisons. Few attempts have been made in face synthesis by taking multi-modal information into account simultaneously. Therefore, the same attribute encoding is also added to pix2pix, termed as MM-pix2pix. Apart from comparison with the baseline algorithms, we also demonstrate the contribution of each part of our MMC-GAN by removing attributes and the local detail generator, separately denoted as w/o attr and w/o detail.

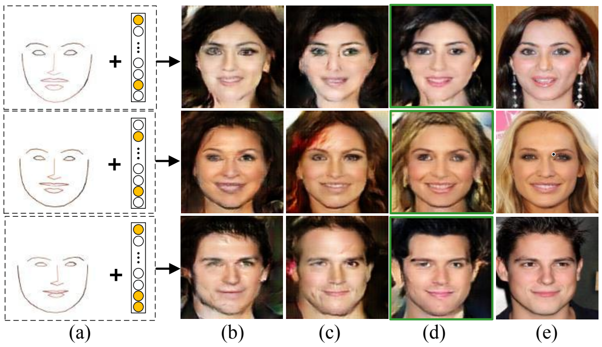

Fig. 2: Sample results of face synthesis, where column (a) is a subject input, and column (b) are synthetic faces by pix2pix [8], and column (c) and (d) are the synthetic faces based on the proposed MMC-GAN without or with the attributes information, respectively. Ground-truth images are shown in (e).

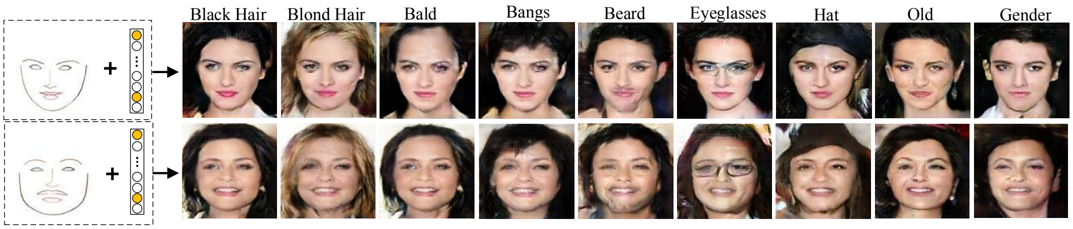

Fig. 3: Synthesis results based on the proposed MMC-GAN conditioned on different attributes. Given sketch and different attributes, we only flip one attribute value for each synthetic face image. The labels are: Black hair, Blond hair, Bald, Bangs, Beard, Eyeglasses, Hat, Old, Gender, respectively.

Table 2: Verification accuracies(%) of the proposed MMC-GAN and compared methods in Sketch-LFWA.

	U-Net	pix2pix	MM-pix2pix	Baseline	MMC-GAN	w/o attr	w/o det
Resnet	70.75	70.93	73.73	58.33	**82.22**	75.98	76.88
VGG-Face	70.22	71.37	75.35	58.70	**82.93**	76.13	77.00

3.2. Face Synthesis

Fig. 2 illustrates some representative examples generated by different methods. It can be seen that pix2pix [8] is able to generate a photo only given a minimal sketch, which is blurry with some defects and missing details. Comparatively, the synthetic images by ”w/o attr” have more convincing details to better reflect global sketch descriptions in the same input, yet failing to control conditional appearance (hair color, skin, etc.). By incorporating attributes with sketch as input, our proposed MMC-GAN model presents plausible face images and completes attribute control, benefiting from the complementary superiority of multi-modal information. The results demonstrate that multi-modal input is capable of capturing facial detailed features as well as realistic state.

Another bonus of the proposed algorithm is that we can perform facial attribute manipulation using MMC-GAN. Given a 19-vector attribute embedding, the results are generated with one attribute value flipped in this attribute vector, as illustrated in Fig. 3. Conditioned on face attributes the generated results are convincing and obvious, which verifies that our model has the capacity to control face synthesis.

3.3. Multi-Modal Recognition

To effectively test our model's identity preserving, we design face verification experiments on Sketch-LFWA. Exsiting VGG-Face model [23] and pre-trained Resnet model on CASIA-WebFace [24] are selected as our baseline models to extract features and then compute verification accuracy with a cosine-distance metric. The matching results on the direct sketch-photo images serve as our baseline results.

Performance evaluation. Tab. 2 and Fig. 4 show the accuracy results and ROC curves. Obviously, our algorithm achieves best performance in verification accuracy and outperforms other alternative methods by a large margin. The main reason is that more face representations are explored in multi-modal input and synthetic uncertainties are also reduced, with a realistic characteristics. Additionally, the performance for MM-pix2pix by jointly utilizing the multi-modal input, although inferior to our algorithm, excels the pix2pix [8], which also demonstrate necessity of multi-conditions. As is shown in Tab.2, it is also observed that the verification accuracy drops apparently without attributes, which means facial attribute information impacts the appearance feature of synthetic image. The performance of MMC-GAN without the local detail generator is also inferior to the approach. Therefore, our multi-modal structure is important for improving performance by our MMC-GAN.

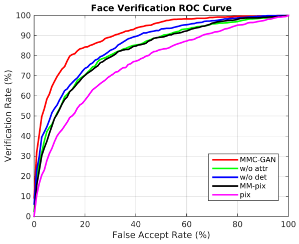

Fig. 4: The ROC curves.

Table 3: Accuracies(%) of the perceptual experiment.

	MMC-GAN
Reco.	89.96 ± 2.51
Veri.	88.80 ± 3.20

Perceptual Evaluation with Humans. We also perform a choice experiment on crowd-sourced workers in order to determine how the generated images can be utilized for the face recognition task in real scenarios. For the first recognition experiment, each worker is presented with 500 synthetic images and each image along with five real images with one image having the same identity, then the workers are asked to pick out the most similar image to the synthetic ones; for the second one, each worker is requested to verify whether given a pair of images belong to same identification. Following the setup, we performed this study across 10 workers and statistics are recorded. The results demonstrate that the generated faces greatly accord with human intuition and are much easier to be recognized for users than the original sketches.

4. CONCLUSION

In this paper, we propose to perform face image generation and recognition by jointly utilizing the multi-modal information provided by minimal sketch and verbal description. Extensive experiments including a perceptual one demonstrate the effectiveness of the proposed MMC-GAN algorithm.

5. REFERENCES

[1] J. Lu, V. E. Liong, and X. Zhou, "Learning compact binary face descriptor for face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 2041–2056, 2015.

[2] M. Kan, S. Shan, and H. Zhang, "Multi-view discriminant analysis," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 188–194, 2016.

[3] D. Gong, Z. Li, and W. Huang, "Heterogeneous face recognition: A common encoding feature discriminant approach," *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2079–2089, 2017.

[4] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[5] B. Xiao, X. Gao, and D. Tao, "Photo-sketch synthesis and recognition based on subspace learning," *Neurocomputing*, vol. 73, no. 4, pp. 840–852, 2010.

[6] Y. Güçlütürk, U. Güçlü, and Rob van Lier, "Convolutional sketch inversion," in *European Conference on Computer Vision*. Springer, 2016, pp. 810–824.

[7] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[8] P. Isola, J.Y. Zhu, and T. Zhou, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[9] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros, "Generative visual manipulation on the natural image manifold," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[10] J.Y. Zhu, T. Park, and P. Isola, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *arXiv preprint arXiv:1703.10593*, 2017.

[11] X. Yan, J. Yang, and K. Sohn, "Attribute2image: Conditional image generation from visual attributes," in *European Conference on Computer Vision*. Springer, 2016, pp. 776–791.

[12] D. P Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4030–4038.

[14] Guim Perarnau, "Invertible conditional gans for image editing," in *NIPS Workshop on Adversarial Training*, 2016.

[15] Y. Sun, Y. Chen, and X. Wang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014, pp. 1988–1996.

[16] G. B Huang, M. Ramesh, and T. Berg, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[17] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[18] R. Huang, S. Zhang, and T. Li, "Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[19] E. S. Gastal and M. M Oliveira, "Domain transform for edge-aware image and video processing," in *ACM Transactions on Graphics (ToG)*. ACM, 2011, vol. 30, p. 69.

[20] D. E King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[21] C. Sagonas, G. Tzimiropoulos, and S. Zafeiriou, "A semi-automatic methodology for facial landmark annotation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 896–903.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[23] O. M Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition.," in *BMVC*, 2015, vol. 1, p. 6.

[24] D. Yi, Z. Lei, and S. Liao, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.