FM^2u-Net: Face Morphological Multi-branch Network for Makeup-invariant Face Verification

Wenxuan Wang[1*‡] Yanwei Fu[1*§] Xuelin Qian[1‡] Yu-Gang Jiang[1†‡] Qi Tian[2¶] Xiangyang Xue[1‡]

[1]Fudan University, China [2]Noah's Ark Lab, Huawei Technologies

Abstract

*It is challenging in learning a makeup-invariant face verification model, due to (1) insufficient makeup/non-makeup face training pairs, (2) the lack of diverse makeup faces, and (3) the significant appearance changes caused by cosmetics. To address these challenges, we propose a unified **Face Morphological Multi-branch Network** (FM^2u-Net) for makeup-invariant face verification, which can simultaneously synthesize many diverse makeup faces through face morphology network (FM-Net) and effectively learn cosmetics-robust face representations using attention-based multi-branch learning network (AttM-Net). For challenges (1) and (2), FM-Net (two stacked auto-encoders) can synthesize realistic makeup face images by transferring specific regions of cosmetics via cycle consistent loss. For challenge (3), AttM-Net, consisting of one global and three local (task-driven on two eyes and mouth) branches, can effectively capture the complementary holistic and detailed information. Unlike DeepID2 which uses simple concatenation fusion, we introduce a heuristic method AttM-FM, attached to AttM-Net, to adaptively weight the features of different branches guided by the holistic information. We conduct extensive experiments on makeup face verification benchmarks (M-501, M-203, and FAM) and general face recognition datasets (LFW and IJB-A). Our framework FM^2u-Net achieves state-of-the-art performances.*

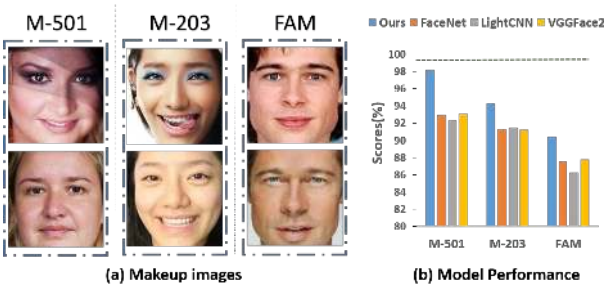


Figure 1. (a) Faces with and without facial cosmetics. The same identity is circled. (b) The bar chart shows the recognition accuracy of the four models on the three makeup face recognition datasets, while the dotted green line shows the average accuracy of the models on the general face recognition dataset LFW [23].

1. Introduction

This paper studies the task of face verification, which judges whether a pair of face images are the same person or not. Despite significant progress has been made by recent deep networks, the general recognition models of such a face task, may not be robust to the makeup faces [15]. On the one hand, we can apparently observe huge visual contrasts of the same identities with and without facial cosmetics in Fig. 1 (a). Furthermore, the perceptual and psychological studies in [38] show that the heavy facial makeup can significantly change facial characteristics, making it challenging and unbelievable to recognize face identities. In addition, we evaluate several popular face recognition models on face recognition datasets, results in Fig. 1 (b) demonstrate a dramatic drop when involving the challenge of makeup. This brings us a natural prospect in learning a robust model for makeup-invariant face verification task.

Recent efforts to address this problem have resorted to synthesizing the non-makeup faces by makeup removal models [28, 7, 29, 13] to help general face models, rather than directly learning a makeup-invariant facial representation as done in this paper. Unfortunately, it is non-trivial to efficiently and effectively learn towards removing makeup, a synthesizer, *e.g.*, Generative Adversarial Net-

*indicates equal contributions.
†indicates corresponding author.
‡Wenxuan Wang, Xuelin Qian, Yu-Gang Jiang, Xiangyang Xue are with School of Computer Science, and Shanghai Key Lab of Intelligent Information Processing, Fudan University. Email: {wxwang19, xlqian15, ygj, xyxue}@fudan.edu.cn.
§Yanwei Fu is with School of Data Science, MOE Frontiers Center for Brain Science, and Shanghai Key Lab of Intelligent Information Processing, Fudan University. Email: yanweifu@fudan.edu.cn.
¶Qi Tian is with Noah's Ark Lab, Huawei Technologies. Email: tian.qi1@huawei.com.

works (GANs). For instance, it is well known that training GANs may suffer from the problem of model collapse. Even worse, the casual nature of makeup faces makes it difficult to learn a good synthesizer. In particular, we highlight the following challenges: (1) *Lack of sufficient paired makeup/non-makeup faces*: It requires prohibitive cost in collecting such large-scale paired makeup face images. (2) *Lack of makeup faces with diverse facial regions*. One may apply various makeup styles to multiple facial regions, *e.g.*, eyelash, nose, cheek, lip, and acne, as shown in Fig. 1 (a). (3) *Huge visual differences caused by cosmetics.* As in Fig. 1 (a), the heavy cosmetics, especially on eyes and mouth regions, greatly degrade the performance of general face recognition models (Fig. 1 (b)).

To this end, this paper presents an end-to-end makeup-invariant face verification model – Face Morphological Multi-branch Network (FM^2u-Net). It is composed of Face Morphology Network (FM-Net) and Attention-based Multi-branch Network (AttM-Net), in addressing the aforementioned three problems. Particularly,

FM-Net. To address the data problem (Challenges (1) and (2)), we propose the FM-Net to synthesize *large-scale* and *diverse* makeup faces. Typically, since heavy cosmetics are usually covered on local patches, *i.e.*, key facial components such as eyes and mouth, it motivates us to synthesize faces directly from these local parts. Specifically, the FM-Net learns a generative model that transfer the key facial components of two inputs via a cycle consistent manner [63]. Functionally, FM-Net stacks two auto-encoders to generate high-quality faces covered by diverse cosmetics, guided by makeup swapping loss, cycle consistent loss, prediction/recognition loss, and ID-preserving loss.

AttM-Net. To achieve a cosmetics-robust model (Challenge (3)) and reduce the significant visual changes on local patches (*e.g.*, eyes in Fig. 1 (a)), the proposed AttM-Net explicitly considers the features of such regions. Specifically, AttM-Net contains four sub-networks to learn the features from the whole face and three facial regions (two eyes and mouth). Furthermore, AttM-Net is learned in an end-to-end manner and we also propose a novel heuristic attention fusion mechanism AttM-FM, which can adaptively weight different parts (global and local) for each particular face.

Contributions. (1) This paper, for the first time, proposes a novel *end-to-end* face morphological multi-branch network (FM^2u-Net) which can simultaneously synthesize large-scale and diverse makeup faces using face morphology network (FM-Net) and effectively learn cosmetics robust face representations through attention-based multi-branch learning network (AttM-Net). (2) FM-Net has two stacked weight-sharing auto-encoders, to synthesize realistic makeup faces by transferring makeup regions of the input faces. (3) AttM-Net contains three local and one global streams to capture the complementary holistic and detailed information. Unlike the simple concatenation fusion used by DeepID2, we propose a new heuristic attention fusion mechanism AttM-FM to adaptively weight the features for one particular face. (4) To thoroughly investigate the problem of learning a makeup-invariant face model, we bring three new datasets to the community, which are rephrased from the several existing datasets: the M-501 [15] dataset is extended by including additional challenging paired makeup/non-makeup images, crawled online. Furthermore, some makeup faces are added to general face recognition datasets (LFW [23] and IJB-A [26]). Extensive experiments are evaluated on these datasets to show that our FM^2u-Net model achieves state-of-the-art performances on these benchmarks.

2. Related Work

Face Recognition. Various deep learning methods have been proposed for general face recognition in the wild, such as FaceNet [40], LightCNN [54], VGGface2 [5], neural tensor networks [19, 21], *etc.*. Apart from that, some works focus on specific challenges of face recognition, such as pose [33, 47, 10], illumination [59, 18], occlusion [53], *etc.*. Unlike the extensive explorations of the aforementioned challenges, less attention is paid to one important problem, cosmetics as shown in Fig. 1. It motivates this work to explore effective solutions on makeuped faces.

Makeup Face Verification. Cosmetics bring enormous challenges for face verification task, due to significant facial appearance changes. Recent works of analyzing makeup faces focus on makeup transfer [46, 14, 39, 32, 6] and makeup recommendation [31, 9, 3, 2]. Few efforts are made on learning a makeup-invariant face verification model. The deep models for general face recognition may not be robust to heavy makeup (*e.g.*, Fig. 1 (b)). To achieve a cosmetics robust face recognition system, Sun *et al.* [45] proposed a model pre-trained on free videos and fine-tuned on small makeup datasets. To alleviate negative effects from makeup, Li *et al.* [29] generated non-makeup images from makeup ones using GANs, and then used the synthesized non-makeup images for recognition. Unlike them, we introduce a unified FM^2u-Net to effectively improve the performance of makeup face verification, which can synthesis many high-quality images with abundant makeup style and extract more cosmetics-robust facial features.

Face Morphology. Recently Sheehan *et al.* [42] suggested that the increased diversity and complexity in human facial morphology are the primary medium of individual identification and recognition. And now face morphology has been used to build a photo-realistic virtual human face [4], face detection [16], 3D face analysis [50], *etc.*. In this work, we aim to generate realistic facial images while keeping the identity information through FM-Net.

Patch-based Face Recognition. While global-based

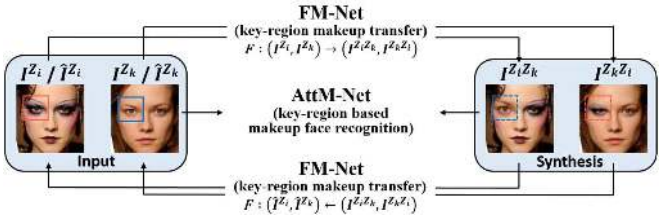


Figure 2. Overview of the proposed FM^2u-Net architecture which contains FM-Net and AttM-Net. FM-Net can synthesize realistic faces with diverse makeup information, and AttM-Net focuses on generating makeup-invariant facial representations. During testing, we directly send the inputs into the trained AttM-Net and use the fused features to perform face verification.

face representation approach prevails, more and more researchers attempt to explore local features [43, 24, 58], which are believed more robust to the variations of facial expression, illumination, and occlusion, *etc.*. For example, at a masquerade party we can identify an acquaintance by the eyes, which are the only visible facial components through mask. Motivated by this, we design AttM-Net to aggregate global and local features by fusion module.

Data Augmentation. Deep models are normally data hungry, therefore, data augmentation has widely been used to increase the amount of training data [27, 36], including flipping, rotating, resizing, *etc.*. Apart from those general data augmentation methods, in the field of face recognition, 3D models [61, 34, 49] and GANs [12, 30, 48, 61] are widely used to synthesizes faces with rich intra-personal variations such as poses, expressions, *etc.*. Note that the idea of synthesizing new images to help recognition has been explored and verified in many tasks, *e.g.*, person re-id [37, 62], and one-shot learning [51, 8]. In this work, we propose a specialized data augmentation for cosmetics-robust face verification. Specifically, we propose FM-Net that can synthesize new faces by swapping the facial components which are usually covered by heavy cosmetics. Unlike [20] which randomly selects the swapping targets in an offline fashion, we choose the swapping targets from similar faces in an end-to-end (online) way via the proposed generative model.

3. Face Morphological Multi-branch Network

Face Verification. Suppose we have the makeup dataset $\mathcal{D}_m = \{(\mathbf{I}_i, z_i)\}_{i=1}^{N}$ of N identities, and a large-scale auxiliary dataset $\mathcal{D}_s = \{(\mathbf{I}_j, z_j)\}_{j=1}^{M}$, *i.e.*, CASIA-WebFace [56]. We use $\mathbf{I}_i$ to indicate the makeup/non-makeup face images of the person z_i. Given a testing face pair $(\mathbf{I}_i, \mathbf{I}_j) \in \mathcal{D}_m$, where one is with cosmetics and the other is not, our goal is to verify whether $z_i = z_j$.

Face Morphology. Normally, heavy makeup is applied to eyes and mouth regions, which greatly affects the accuracy of face verification. Thus we focus on the key facial patches (left eye, right eye, and mouth), which are generated by the 32×32, 32×32 and 16×48 bounding boxes centered at three landmarks via MTCNN [57] respectively. Here, we simplify the symbol of face image $(\mathbf{I}_i, z_i)$ as $\mathbf{I}^{z_i}$, and denote $\mathbf{I}^{z_i z_j}$ as a face morphological image with identity z_i, where makeups in the key patch $\mathbf{P}^{z_j}$ of $\mathbf{I}^{z_j}$ is transferred to the corresponding patch $\mathbf{P}^{z_i}$ of $\mathbf{I}^{z_i}$. Note that we can produce three different variations of $\mathbf{I}^{z_i z_j}$ since there are three handpicked key patches. In this work, we use $\mathbf{I}^{z_i z_j}$ to refer to any of them in general.

Our Framework. To tackle the data problem and achieve makeup-invariant face recognition, we propose a unified face morphological multi-branch network (FM^2u-Net), which includes two modules: face morphology network (FM-Net), and attention-based multi-branch network (AttM-Net). The FM-Net can conduct the face morphological operations to synthesize some realistic faces covered with diverse cosmetics; and AttM-Net, which consists of four (one global and three local) branches, can effectively capture the discriminative local and global features and then fuse them to generate a cosmetics-robust representation guided by the global branch through feature fusion module. The whole framework is shown in Fig. 2, and it is trained in an end-to-end manner by optimizing the loss functions of both modules,

$$\{\Omega, \Theta\} = \underset{\Omega,\Theta}{\operatorname{argmin}} \mathcal{L}_{\mathrm{FM-Net}}(\Omega) + \lambda \mathcal{L}_{\mathrm{AttM-Net}}(\Theta) \quad (1)$$

where $\mathcal{L}_{\mathrm{FM-Net}}(\Omega)$ and $\mathcal{L}_{\mathrm{AttM-Net}}(\Theta)$ are the losses for FM-Net and AttM-Net respectively; Ω and Θ indicate the corresponding parameters; λ is a trade-off hyper-parameter.

3.1. Face Morphology Network

As introduced in Sec. 1, the data problem (lack of sufficient and diverse makeup/non-makeup paired training data) makes the cosmetics-robust face recognition very challenging. It is observed that the cosmetics are normally covered in three local facial regions (two eyes and mouth) in Fig. 1 (a). It motivates us to synthesize abundant, diverse and realistic makeup faces by transferring the makeup patches (two eyes and mouth) between two similar faces. The images generated in such way can not only keep the identity information of most facial regions, but also increase the diversity of makeup information by introducing the local patches with cosmetics. To achieve this, we naturally turn to generative models. However, there are no ground truths of realistic faces with such swapped facial components, to guide this generating process. Thus we use the sets of original images and facial patches as supervision information, and employ cycle consistent loss [63] to guide realistic makeup face generation. In this work, we propose a generative model, Face Morphology Network (FM-Net), to achieve this.

FM-Net is stacked by two weight-sharing auto-encoders. We summarize the data flow of FM-Net: for the image

$\mathbf{I}^{z_i} \in \mathcal{D}_m$, we compute its most similar Top-K facial images $\mathbf{I}^{z_k} \in \mathcal{D}$[1], where $\mathcal{D} = \mathcal{D}_m \cup \mathcal{D}_s$, by comparing the similarity between the features extracted from the face recognition model $\phi(\cdot)$, which is the LightCNN-29v2 [54, 1] pretrained on $\mathcal{D}_s$ here.

As illustrated in Fig. 2, given an input image $\mathbf{I}^{z_i}$, its corresponding Top-K image $\mathbf{I}^{z_k}$ and their key patch locations, the first auto-encoder $\mathcal{F}$ learns the mapping as face morphology operation [2] $(\mathbf{I}^{z_i}, \mathbf{I}^{z_k}) \rightarrow (\mathbf{I}^{z_i z_k}, \mathbf{I}^{z_k z_i})$. Specifically, $(\mathbf{I}^{z_i z_k}, \mathbf{I}^{z_k z_i})$ is generated from $(\mathbf{I}^{z_i}, \mathbf{I}^{z_k})$ by swapping the cosmetics between patches of $\mathbf{P}^{z_i}$ and $\mathbf{P}^{z_k}$, and we use $\mathbf{P}^{z_i z_k}$ and $\mathbf{P}^{z_k z_i}$ to denote the transferred patches in $\mathbf{I}^{z_i z_k}$ and $\mathbf{I}^{z_k z_i}$, respectively. Motivated by cycle consistent loss, we reconstruct the original faces using the second auto-encoder (whose weights are shared with the first one) via the same projection $\mathcal{F}: (\mathbf{I}^{z_i z_k}, \mathbf{I}^{z_k z_i}) \rightarrow \left(\hat{\mathbf{I}}^{z_i}, \hat{\mathbf{I}}^{z_k}\right)$, where $\hat{\mathbf{I}}^{z_i}$ and $\hat{\mathbf{I}}^{z_k}$ are the reconstructed version of the original images $\mathbf{I}^{z_i}$ and $\mathbf{I}^{z_k}$, and we let $\hat{\mathbf{P}}^{z_i}$ and $\hat{\mathbf{P}}^{z_k}$ to indicate the reconstructed regions in $\hat{\mathbf{I}}^{z_i}$ and $\hat{\mathbf{I}}^{z_k}$, respectively. Clearly, the intermediate generated paired faces $(\mathbf{I}^{z_i z_k}, \mathbf{I}^{z_k z_i})$ are the synthetic faces with one of facial makeups transferred. These synthetic realistic faces can greatly enlarge training data and introduce more diverse makeup changes.

To synthesize realistic faces, FM-Net consists makeup swapping loss, cycle consistent loss, prediction/recognition loss, ID-preserving loss, and a regularization term.

Makeup Swapping Loss. The makeup swapping loss constrains the local patch swapping process across face images. To accurately introduce the abundance of makeup information to the facial parts, it encourages the selected patches to stay the same among the swapping process:

$$\mathcal{L}_{patch} = \sum_{i,k} \left(|\mathbf{P}^{z_i z_k} - \mathbf{P}^{z_k}| + |\mathbf{P}^{z_k z_i} - \mathbf{P}^{z_i}| + \left|\hat{\mathbf{P}}^{z_i} - \mathbf{P}^{z_k z_i}\right| + \left|\hat{\mathbf{P}}^{z_k} - \mathbf{P}^{z_i z_k}\right| \right) \quad (2)$$

Cycle Consistent Loss. Aiming to remain the generated faces keeping original identity features, we constrain the reconstructed faces from the second auto-encoder as similar as inputs, so we use the cycle consistent loss as constraint,

$$\mathcal{L}_{cycle} = \sum_{i,k} \left(\left|\mathbf{I}^{z_i} - \hat{\mathbf{I}}^{z_i}\right| + \left|\mathbf{I}^{z_k} - \hat{\mathbf{I}}^{z_k}\right| \right) \quad (3)$$

Prediction/Recognition loss. We here use cross-entropy loss for face recognition to learn discriminative and identity-sensitive features, which supervise the identity of generated images:

[1]For simplicity, we use $\mathbf{I}^{z_k}$ to refer to any of Top-K images in general.

[2]To ease understanding, we omit the symbols of the input coordinates for patches.

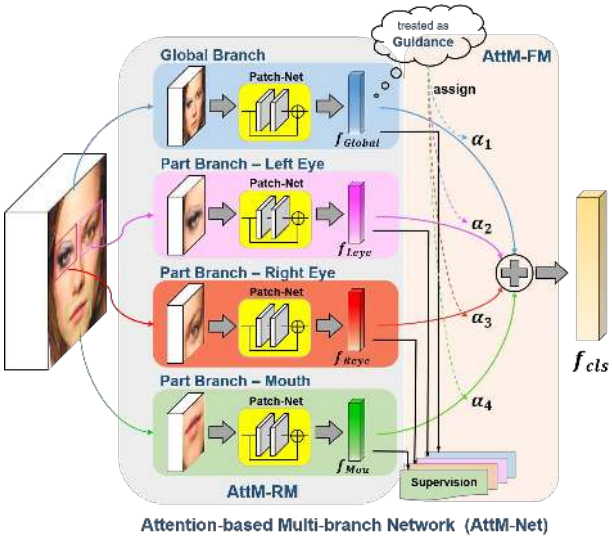


Figure 3. AttM-Net learns the makeup-invariant and identity-sensitive features, which contains AttM-RM and AttM-FM. To obtain cosmetics-robust facial representations, AttM-RM focuses on the facial patches often wearing makeup and AttM-FM discovers the more discriminating channels to generate final fusion feature.

$$\mathcal{L}_{pre} = -\frac{1}{N} \sum \left[z_i \log \left(\frac{e^{\tilde{\phi}(\mathbf{I}^{z_i z_k})}}{\sum e^{\tilde{\phi}(\mathbf{I}^{z_i z_k})}} \right) + z_k \log \left(\frac{e^{\tilde{\phi}(\mathbf{I}^{z_k z_i})}}{\sum e^{\tilde{\phi}(\mathbf{I}^{z_k z_i})}} \right) \right] \quad (4)$$

where $\tilde{\phi}(\cdot)$ indicates the feature extractor $\phi(\cdot)$ followed by one FC layer for classifier.

ID-preserving Loss. It is important to make the generated faces keep the identity information, thus we require the identity prediction of the generated image to be consistent with the original face. Inspired by the perceptual loss [25] in the image generation task, we use the ID-preserving loss:

$$\mathcal{L}_{id} = |\phi(\mathbf{I}^{z_i}) - \phi(\mathbf{I}^{z_i z_k})| + |\phi(\mathbf{I}^{z_k}) - \phi(\mathbf{I}^{z_k z_i})| \quad (5)$$

FM-Net Loss. Combining the aforementioned losses, FM-Net loss is defined as:

$$\mathcal{L}_{\text{FM-Net}}(\Omega) = \mu_1 \mathcal{L}_{patch}(\Omega) + \mu_2 \mathcal{L}_{cycle}(\Omega) + \mu_3 \mathcal{L}_{pre}(\Omega) + \mu_4 \mathcal{L}_{id}(\Omega) + \mu_5 P(\Omega) \quad (6)$$

where $\{\mu_1, \mu_2, \mu_3, \mu_4, \mu_5\}$ are the weights, $P(\cdot)$ indicates the L_2 regularization term.

3.2. Attention-based Multi-branch Network

Since cosmetics can significantly change facial appearance, it is difficult to learn cosmetics robust face representations, where seriously degrades the performance of makeup-invariant face verification. It is observed that heavy cosmetics are mainly applied to key local facial patches (two eyes and mouth). This motivates us in learning cosmetics-robust identity features on these local patches. To

achieve this, in this work, we propose the AttM-Net, consisting of an attention-based multi-branch recognition module (AttM-RM) and a feature fusion module (AttM-FM), as in Fig. 3. The AttM-Net contains four networks to extract one global and three local (tow eyes and mouth) features, and AttM-FM fuses these feautres under the guidance learned from the global one, making it possible to dynamically weight the contributions of four features to the final verification decision.

Attention-based Multi-branch Recognition Module. It is not new to use both global and local features for face recognition problem. One representative method is DeepID2 [44] which works on general face recognition and achieves impressive performance using both global faces and local patches. However, DeepID2 does not have clear guidance on which local patches are discriminative. Thus DeepID2 randomly selects 200 local patches for training, which is a quite brutal force. To achieve cosmetics-robust face verification, clearly, the cosmetics are mainly covered on three key facial components (two eyes and mouth), motivating us to focus on these three patches. Therefore, AttM-RM has four networks of Patch-Net which all use LightCNN-29v2 [54] as the backbone shown in Fig. 3. These four networks generate one global feature (f_{Global}) and three local features (f_{Leye}, f_{Reye} and f_{Mou}).

Feature Fusion Module. Unlike DeepID2 which uses simple concatenation for feature fusion, our AttM-FM performs the fusion of the features from those four streams in a heuristic way (attention). Different from previous works [36, 55] as well, we employ the global features f_{Global} to guide the refinement of features from other local branches, which encourages them to discover the channels containing more discriminative and makeup-invariant features. Specially, we first utilize the features from the global branch to compute weights $\alpha = \varphi(\mathbf{W} \cdot f_{Global})$, where $\alpha = \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\} \in \mathbb{R}^{4\times 1}$, $W \in \mathbb{R}^{4\times C\times H\times W}$ is the weight of the convolution layers with 1×1 kernels; here C, H and W represent the input feature channels, height and width respectively. $\varphi(\cdot)$ indicates the activation function, *i.e.*, ReLU. Then the fused feature f_{cls} is the concatenation of weighted four features:

$$f_{cls} = [\alpha_1 f_{Leye} \ ; \ \alpha_2 f_{Reye} \ ; \ \alpha_3 f_{Mou} \ ; \ \alpha_4 f_{Global}] \quad (7)$$

Face Recognition Loss. Since this work is finally for face recognition, the cross-entropy loss is performed on the four features (from four streams) and the fused one f_{cls}:

$$\mathcal{L}_{\mathrm{AttM-Net}} = \gamma_1\mathcal{L}_{leye} + \gamma_2\mathcal{L}_{reye} + \gamma_3\mathcal{L}_{mou} + \gamma_4\mathcal{L}_{global} + \gamma_5\mathcal{L}_{cls} \quad (8)$$

where $\mathcal{L}_{leye}$, $\mathcal{L}_{reye}$, $\mathcal{L}_{mou}$ and $\mathcal{L}_{global}$ are the face recognition loss functions for four streams, and $\mathcal{L}_{cls}$ is classification loss working on the fused feature f_{cls} from the final layer of the AttM-Net. $\{\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5\}$ are the weights. For the inference process, given one aligned test face pair, we directly apply AttM-Net to extract image features, and use the fused feature f_{cls} to compute the similarity score.

Figure 4. Sample image pairs of the extended makeup dataset. The first row images are the faces without makeup, the second row images are the corresponding makeup faces of the same person. In addition to makeup, this dataset may also have the facial images of large areas of acne, glasses occlusion, head posture changes and so on. The same person is circled by a blue dash line.

4. Experiment

4.1. Datasets and Settings

Makeup Dataset. (1) M-501: Guo *et al.* [15] proposed a makeup face database containing 1,002 face images (501 pairs). (2) M-203: This dataset collected in [45] contains 203 pairs corresponding to a female individual. (3) FAM: Face Makeup (FAM) Database [22] involves 519 subjects, 222 of them are male and the remaining 297 are female. (4) Extended makeup dataset (Extended): We triple the test data of M-501 by collecting more paired makeup/non-makeup faces from the Internet as shown in Fig. 4. Note that we only enlarge the test set while the training set stays the same. Each subject has two images in the above datasets: one is with makeup and the other is not.

Extended general face datasets. We introduce LFW+ and IJB-A+ datasets which are the extension of the general face datasets (LFW [23] and IJB-A [26]) by adding makeup face images. (1) LFW+: This dataset consists of original LFW (13,233 faces) and the 2,726 testing images from four makeup dataset. Thus, we test 10 splits in total, and each split includes 600 pairs from LFW and makeup datasets, respectively (total 1200 pairs); each split has half positive and half negative pairs. (2) IJB-A+: IJB-A dataset contains 5712 images and 2085 videos from 500 subjects, with an average of 11.4 images and 4.2 videos per subject. We follow the original IJB-A testing standards and extend the testing data with four makeup testing data.

Implementation details. We use Pytorch for Implementation, and utilize SGD to train our models. Dropout is applied after the last fully connected layer with the ratio of 0.5. The input images are resized to 144×144 firstly, then randomly cropped to 128×128. We apply face morphological operation to extend the training data with $K = 20$. The initial learning rate of fully connected layers is set as 0.01 and other layers as 0.001, they are gradually decreased to zero from the 80th epoch, and stopped after the 150th epoch. We

	Method	M-501	M-203	FAM	Ext
FR#	CSML [35]	–	–	62.40	–
	FaceNet [40]	92.92	91.32	87.51	90.26
	MTNet [52]	92.78	91.47	85.22	89.39
	VGGFace2 [5]	93.16	91.23	87.81	91.92
	LightCNN [54]	94.21	91.50	86.30	90.03
	ArcFace [11]	93.67	92.39	88.52	92.17
	AdaCos [60]	92.89	92.41	88.73	91.92
MFR*	CorrNet [22]	–	–	59.60	–
	PCA+PLS [15]	80.50	–	–	–
	TripletNet [45]	82.40	68.00	–	–
	BLAN [29]	94.80	92.30	88.10	–
Our FM^2u-Net		**98.12**	**94.34**	**90.43**	**96.56**

Table 1. Results on four Makeup datasets. Ext represents the extended makeup dataset. FR#: general face recognition (FR) models. MFR*: makeup face recognition (MFR) models.

set the mini-batch size as 64, λ=1, μ_1=5, μ_2=1, $\mu_3 = 1$, $\mu_4 = 2$ and μ_5= 5e^{-4}, γ_1, γ_2, γ_3, γ_4 are equally as 1 and γ_5 as 2. Our model is trained by one NVIDIA GeForce GTX 1080Ti GPU. On M-501 and FAM, our models converge with 140 epochs (around 15 hours), and converges at 130th epoch on M-203 (about 12 hours).

Evaluation metrics. (1) Makeup face verification task: we use five-fold cross validation for our experiments as [15, 45, 22]. In each round, there are 4/5 individuals or face pairs in the training set, and the remaining 1/5 for testing, and there is no overlapping between training and test sets. All the positive pairs are involved during the test and the same number of pairs of negative samples is randomly selected. For example, in the M-501 dataset, there are about 200 paired faces for testing each round, including positive and negative pairs. The averaged results are reported over five rounds to measure the performance of algorithms. (2) General face verification task: We use ten-split evaluations with standard protocols as [5, 26], where we directly extract the features from the models for the test sets and use cosine similarity score. For LFW+, we test our algorithm on 12000 face pairs and mean accuracy is reported. In the IJB-A+ dataset, for 1:1 face verification, the performance is reported using the true accept rates (TAR) vs. false positive rates (FAR), and the performance is reported using the Rank-N as the metrics for 1:N face identification.

4.2. Results on Makeup Face Recognition Datasets

We evaluate several competitors including general and makeup-based face recogntion methods on four makeup datasets. All results are shwon in Tab. 1.

Comparisons with general face recognition methods. To make a fair comparison, we re-train all the compared general FR models with the same data (including synthetic makeup data from FM-Net). Compared with existing methods, FM^2u-Net model gets 96.56% on the extended makeup dataset, which has remarkable verification accuracy improvement. FM^2u-Net model is clearly better than those widely used general face recognition networks, indicating that AttM-Net is a more effective way of learning makeup-invariant face features adaptively fused by four branches.

Comparisons with the state-of-the-art. As shown in Tab. 1, FM^2u-Net achieves the verification accuracy of 98.12% on M-501, showing 3.32% and 15.72% improvement over BLAN [29] and TripletNet [45] respectively. In addition, FM^2u-Net significantly outperforms BLAN: 94.34% vs. 92.30% on M-203 and 90.43% vs. 88.10% on FAM. Compared with the BLAN, which uses GANs to remove the cosmetics from makeup faces, we enrich the makeup training samples through the swapping of local regions, and let the network generate discriminative features learning from the parts often with heavy cosmetics.

Figure 5. Samples of the synthesized images with makeup and non-makeup. The same person is identified by blue dash circles.

4.3. Results on General Face Recognition Datasets

It is important to evaluate our method on general face recognition datasets. We compare FM^2u-Net with some mainstream models built for general face recognition on LFW+ and IJB-A+ datasets. From Tab. 2, our model achieves the highest face verification accuracy of 97.89% on LFW+. FM^2u-Net can also significantly outperform all existing methods on IJB-A+ for both 1:1 verification and 1:N identification tasks. Such experimental results show the strong generalization capability of our method and usefulness in real applications. The superior performance results from (1) the realistic big diverse makeup data generated by FM-Net and (2) the strong feature learning capacity of AttM-Net.

4.4. Ablation Study

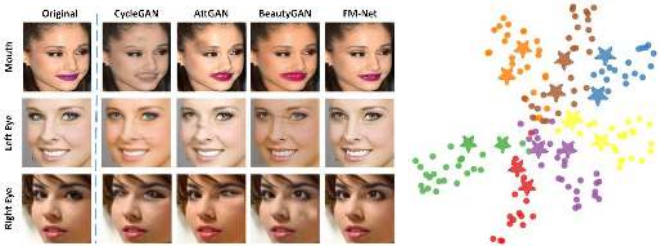
4.4.1 Qualitative Results

Quality of synthesized images. (1) Figure 5 shows the synthetic images by our model. We can see that these images introduce diversity to the local patches, while the remaining areas are kept the same as original input faces. Note that the synthetic faces are realistic and the transferred makeup facial areas are very smooth. (2) To evaluate the quality of our synthesized images, we compare FM-Net with other generative models, such as CycleGAN [63], AttGAN [17], Beau-

Method	LFW+	IJB-A+ (1:1 Verification)			IJB-A+ (1:N Identification)	
		FAR=0.001	FAR=0.01	FAR=0.1	Rank-1	Rank-5
FaceNet [40]	93.87	85.2 ± 1.1	88.5 ± 0.5	89.5 ± 0.2	91.0 ± 0.6	94.9 ± 0.1
MTNet [52]	91.11	81.4 ± 1.8	85.7 ± 0.6	87.7 ± 0.2	83.7 ± 0.7	85.8 ± 0.2
VGGFace2 [5]	93.21	85.1 ± 1.3	88.9 ± 0.6	90.8 ± 0.1	92.5 ± 0.3	94.9 ± 0.1
LightCNN [54]	92.56	85.7 ± 1.2	88.1 ± 0.8	89.4 ± 0.1	90.4 ± 0.3	93.2 ± 0.2
ArcFace [11]	94.98	86.0 ± 1.5	89.8 ± 0.7	91.2 ± 0.2	92.9 ± 0.5	95.6 ± 0.2
AdaCos [60]	95.12	86.3 ± 1.3	89.7 ± 0.9	91.1 ± 0.1	93.1 ± 0.3	95.3 ± 0.1
Our FM^2u-Net	**97.89**	**88.4 ± 1.5**	**91.2 ± 1.1**	**92.2 ± 0.2**	**94.1 ± 0.2**	**96.5 ± 0.1**

Table 2. Results on general face recognition datasets.

tyGAN [28], As shown in Fig. 6 (a), compared with FM-Net, there are more incur artifacts (noise, deformed parts) in images generated from other models, which are not necessarily good for face recognition. (3) For face recognition, we expect the synthetic faces not only look realistic but preserve the identity information. Fig. 6 (b) visualizes the distributions of original and synthetic data using sample images in the M-501 dataset via t-SNE. One color represents one identity. We can see the synthetic data are clustered around the original images with the same identities. It means our generation method can effectively preserve the identity information, which is essential to train a face recognition model.



(a) (b)

Figure 6. (a) Comparison with other generative models. (b) Visualization of 7 paired makeup/non-makeup images (drawn as stars) and synthesized images (drawn as dots) using t-SNE. One color indicates one identity.

Impact of loss functions of FM-Net. To synthesize realistic makeup faces, FM-Net uses four key losses: (a) component swapping loss (patch loss), (b) cycle consistent loss, (c) prediction loss and (d) ID-preserving loss as detailed in Sec. 3.1. To verify the impact of these losses, we remove these losses respectively during training and present the results in Fig. 7. Without loss (a), the patch swapping is not conducted at all, showing that loss (a) is the essential one for swapping patch. Removing loss (b), (c), (d), the quality of synthetic faces all degrades to different extends. In particular, removing loss (b), the synthetic faces become very blurry. It shows the usefulness of all the proposed losses.

Interpretation of the learned models. Grad-CAM [41] is a good tool to interpret the learned CNN models via visualization. Given the learned CNNs, it can visualize the discriminative areas of the test images. In Fig. 8, we visualize a group of faces with makeup based on the learned features of global branch in AttM-Net model and baseline model (the pre-trained LightCNN-29v2 model fine-tuned on makeup data). Compared with the baseline model, our global branch focuses more on the facial areas with less makeup, such as cheeks, nose, forehead, *etc*., rather than the areas (eyes and mouth) with heavy makeup. It means we can find the discriminative area and ignore those confusing areas (heavy makeup areas), thus our model successfully learns the makeup-robust facial representations.

Figure 7. Ablation study for the losses of FM-Net. The above generated images are transferred the left eyes according to the inputs.

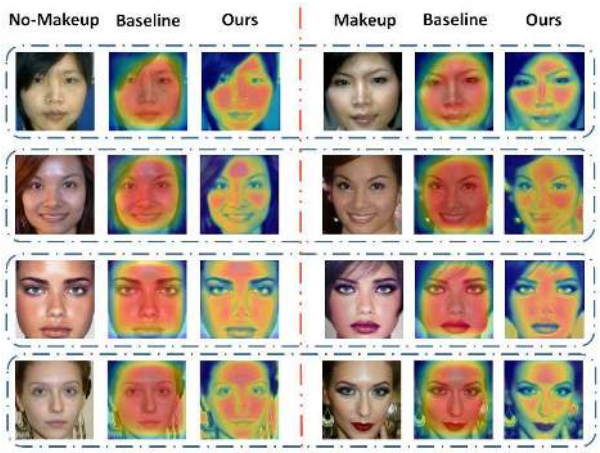


Figure 8. The visualization of discriminative areas of baseline model and the global branch of FM^2u-Net using Grad-CAM [41].

4.4.2 Quantitative Results

The efficacy of different modules in FM^2u-Net. In Tab. 3, we list three variants of FM^2u-Net: 'Baseline': use the makeup

Method	M-501	LFW+
Baseline	94.21	94.01
w.o. FM-Net	95.22	95.34
w.o. AttM-Net	95.98	96.07
Our FM^2u-Net	**98.12**	**97.89**

Table 3. Results of ablation study on two datasets. 'w.o.' means removing this part module from the whole framework FM^2u-Net.

data to fine-tune LightCNN-29v2 model which is pre-trained on CASIA-WebFace; 'Without (w.o.) FM-Net': train AttM-Net on the original training data without synthetic images; 'Without (w.o.) AttM-Net': use LightCNN-29v2 to learn the features from FM-Net outputs. we compare the performance of the aforementioned variants of our model on a makeup (M-501) and a general face verification (LFW+) dataset. It is obvious that FM^2u-Net greatly outperforms 'w.o. FM-Net' and 'w.o. AttM-Net' in Tab. 3: 98.12% vs. 95.22% and 98.12% vs. 95.98% on M-501, showing the effectiveness of FM-Net and AttM-Net.

Method		M-501	LFW+
FM-Net Variants	Hard Replacement	97.14	95.91
	Random Noise Patch	96.48	96.24
	Random Patch	96.92	96.84
Other Generative Models	CycleGAN [63]	87.89	92.12
	AttGAN [17]	93.46	94.01
	BeautyGAN [28]	91.38	94.24
	BeautyGlow [7]	94.07	95.90
	LADN [13]	92.36	94.62
Our FM^2u-Net		**98.12**	**97.89**

Table 4. Results of the variants of FM-Net in FM^2u-Net and results compared with other generative models.

The efficacy of FM-Net. To further study FM-Net, we also compare it with the variants 'Hard Replacement' (we do not use the auto-encoder to *softly* generate the new makeup faces, instead, we swap facial components in a *hard* copy-and-paste way among face images.), 'Random Noise Patch' (expand the training data by adding random noise to the three key regions) and 'Random patch' (using FM-Net, we randomly choose local patches to swap rather than the three handpicked key patches (two eyes and mouth)). In Tab. 4, compared with 'Hard Replacement' and 'Random Noise Patch', our method achieves superior performance, meaning that FM-Net generated faces can capture more diversity and preserve the ID information. Using 'Random Patch' leads to a performance decrease, which demonstrates that the careful chosen facial regions are very discriminative.

It is well known that the synthesized images by GANs are generally not helpful to recognition, despite impressive visual results achieved by GANs. We also conduct some quantitative experiments compared with other generative models, such as, CycleGAN [63], AttGAN [17], BeautyGAN [28], BeautyGlow [7], and LADN [13]. For fairness, we utilize AttM-Net to extract facial features for all generative models. The results in Tab. 4 confirms that FM-Net can generate higher quality training data, and such swapping among facial patches introduces more diverse makeup changes which is useful to learn makeup-invariant features.

Method		M-501	LFW+
Single Branch	Left_Eye	70.31	67.55
	Right_Eye	72.49	68.93
	Mouth	66.17	61.28
	Global Face	96.03	96.24
Fusion	Concatenation	97.44	97.20
	Average	96.65	97.18
	Score	97.03	96.81
Our FM^2u-Net		**98.12**	**97.89**

Table 5. Results of the variants of AttM-Net in FM^2u-Net.

The effectiveness of AttM-Net. We conduct the individual network evaluation ('Single Branch': we train four networks independently and evaluate the performance of each network (Left_Eye/Right_eye/Mouth/Global Face Verification)) to demonstrate that the features from four branches in AttM-Net are complementary to each other. In Tab. 5, the global feature works better than other local features because global features can capture more information, and two eyes are more discriminative than the mouth area. To verify the effectiveness of AttM-FM fusing the features from different branches, we compare with other widely used fusion methods: concatenation fusion, average fusion, and score fusion. We can see that our FM^2u-Net (attention fusion) outperforms all the competing fusion methods. It is also observed that the fusion of four networks in any fusion method works better than the individual network, showing the strong complementarity of the four networks/patches.

5. Conclusions

This paper proposes FM^2u-Net to learn makeup-invariant face representation. FM^2u-Net contains FM-Net and AttM-Net. FM-Net can effectively synthesize many diverse makeup faces, and AttM-Net can capture the complementary global and local information. Besides, AttM-Net applies AttM-FM to adaptively fuse the features from the different branches. Extensive experiments are conducted and results show our method can achieve competitive performance on makeup and general face recognition benchmarks. We also do ablation studies to verify the efficacy of each component in our model.

6. Acknowledgement

This work was supported in part by NSFC Projects (U1611461,61702108) , Science and Technology Commission of Shanghai Municipality Projects (19511120700, 19ZR1471800), Shanghai Municipal Science and Technology Major Project (2018SHZDZX01), and Shanghai Research and Innovation Functional Program (17DZ2260900).

References

[1] https://github.com/AlfredXiangWu/LightCNN.

[2] Taleb Alashkar, Songyao Jiang, and Yun Fu. Rule-based facial makeup recommendation system. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 325–330. IEEE, 2017.

[3] Taleb Alashkar, Songyao Jiang, Shuyang Wang, and Yun Fu. Examples-rules guided deep neural network for makeup recommendation. In *AAAI*, pages 941–947, 2017.

[4] AF Ayoub, Y Xiao, B Khambay, JP Siebert, and D Hadley. Towards building a photo-realistic virtual human face for craniomaxillofacial diagnosis and treatment planning. *International journal of oral and maxillofacial surgery*, 36(5):423–428, 2007.

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[6] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Hung-Jen Chen, Ka-Ming Hui, Szu-Yu Wang, Li-Wu Tsao, Hong-Han Shuai, and Wen-Huang Cheng. Beautyglow: On-demand makeup transfer framework with reversible generative network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10042–10050, 2019.

[8] Zitian Chen, Yanwei Fu, Yu-Xiong Wang, Lin Ma, Wei Liu, and Martial Hebert. Image deformation meta-networks for one-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8680–8689, 2019.

[9] Kyung-Yong Chung. Effect of facial makeup style recommendation on visual sensibility. *Multimedia Tools and Applications*, 71(2):843–853, 2014.

[10] Jiankang Deng, Shiyang Cheng, Niannan Xue, Yuxiang Zhou, and Stefanos Zafeiriou. Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[13] Qiao Gu, Guanzhi Wang, Mang Tik Chiu, Yu-Wing Tai, and Chi-Keung Tang. Ladn: Local adversarial disentangling network for facial makeup and de-makeup. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10481–10490, 2019.

[14] Dong Guo and Terence Sim. Digital face makeup by example. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 73–79. IEEE, 2009.

[15] Guodong Guo, Lingyun Wen, and Shuicheng Yan. Face authentication with makeup changes. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):814–825, 2014.

[16] Chin-Chuan Han, Hong-Yuan Mark Liao, Gwo-Jong Yu, and Liang-Hua Chen. Fast face detection via morphology-based pre-processing. *Pattern Recognition*, 33(10):1701–1712, 2000.

[17] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 2019.

[18] Guosheng Hu, Chi Ho Chan, Fei Yan, William Christmas, and Josef Kittler. Robust face recognition by an albedo based 3d morphable model. In *IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[19] Guosheng Hu, Yang Hua, Yang Yuan, Zhihong Zhang, Zheng Lu, Sankha S Mukherjee, Timothy M Hospedales, Neil M Robertson, and Yongxin Yang. Attribute-enhanced face recognition with neural tensor fusion networks. In *ICCV*, pages 3744–3753, 2017.

[20] Guosheng Hu, Xiaojiang Peng, Yongxin Yang, Timothy M Hospedales, and Jakob Verbeek. Frankenstein: Learning deep face representations using small data. *IEEE Transactions on Image Processing*, 27(1):293–303, 2018.

[21] Guosheng Hu, Yongxin Yang, Dong Yi, Josef Kittler, William Christmas, Stan Z Li, and Timothy Hospedales. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 142–150, 2015.

[22] Junlin Hu, Yongxin Ge, Jiwen Lu, and Xin Feng. Makeup-robust face verification. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2342–2346. IEEE, 2013.

[23] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2007.

[24] Tai-Xiang Jiang, Ting-Zhu Huang, Xi-Le Zhao, and Tian-Hui Ma. Patch-based principal component analysis for face recognition. *Computational intelligence and neuroscience*, 2017, 2017.

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[26] Brendan F Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, and Anil K Jain. Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1931–1939, 2015.

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[28] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 645–653. ACM, 2018.

[29] Yi Li, Lingxiao Song, Xiang Wu, Ran He, and Tieniu Tan. Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification. *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[30] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3362–3371, 2017.

[31] Luoqi Liu, Junliang Xing, Si Liu, Hui Xu, Xi Zhou, and Shuicheng Yan. Wow! you are so beautiful today! *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(1s):20, 2014.

[32] Si Liu, Xinyu Ou, Ruihe Qian, Wei Wang, and Xiaochun Cao. Makeup like a superstar: Deep localized makeup transfer network. *arXiv preprint arXiv:1604.07102*, 2016.

[33] Iacopo Masi, Stephen Rawls, Gerard Medioni, and Prem Natarajan. Pose-aware face recognition in the wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[34] Iacopo Masi, Anh Tuáº¥n Tráº§n, Tal Hassner, Jatuporn Toy Leksut, and Gérard Medioni. Do we really need to collect millions of faces for effective face recognition? In *European Conference on Computer Vision*, pages 579–596. Springer, 2016.

[35] Hieu V Nguyen and Li Bai. Cosine similarity metric learning for face verification. In *Asian conference on computer vision*, pages 709–720. Springer, 2010.

[36] Xuelin Qian, Yanwei Fu, Yu-Gang Jiang, Xiangyang Xue, and Tao Xiang. Multi-scale deep learning architectures for person re-identification. ICCV, 2017.

[37] Xuelin Qian, Yanwei Fu, Wenxuan Wang, Tao Xiang, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. *ECCV*, 2018.

[38] Gillian Rhodes, Alex Sumich, and Graham Byatt. Are average facial configurations attractive only because of their symmetry? *Psychological Science*, 10(1):52–58, 1999.

[39] Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel. Computer-suggested facial makeup. In *Computer Graphics Forum*, volume 30, pages 485–492. Wiley Online Library, 2011.

[40] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[41] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[42] Michael J Sheehan and Michael W Nachman. Morphological and population genomic evidence that human faces have evolved to signal individual identity. *Nature communications*, 5:4800, 2014.

[43] Yu Su, Shiguang Shan, Xilin Chen, and Wen Gao. Hierarchical ensemble of global and local classifiers for face recognition. *IEEE Transactions on image processing*, 18(8):1885–1896, 2009.

[44] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014.

[45] Yao Sun, Lejian Ren, Zhen Wei, Bin Liu, Yanlong Zhai, and Si Liu. A weakly supervised method for makeup-invariant face verification. *Pattern Recognition*, 66:153–159, 2017.

[46] Wai-Shun Tong, Chi-Keung Tang, Michael S Brown, and Ying-Qing Xu. Example-based cosmetic transfer. In *Computer Graphics and Applications, 2007. PG'07. 15th Pacific Conference on*, pages 211–218. IEEE, 2007.

[47] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[48] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *CVPR*, volume 3, page 7, 2017.

[49] Anh Tuáº¥n Tráº§n, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018.

[50] Enrico Vezzetti and Federica Marcolin. Geometry-based 3d face morphology analysis: soft-tissue landmark formalization. *Multimedia tools and applications*, 68(3):895–929, 2014.

[51] Yuxiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. *CVPR*, 2018.

[52] Zhanxiong Wang, Keke He, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *ICMR*. ACM, 2017.

[53] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

[54] Xiang Wu, Ran He, Zhenan Sun, and Tieniu Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, pages 2884–2896, 2018.

[55] Fen Xiao, Wenzheng Deng, Liangchan Peng, Chunhong Cao, Kai Hu, and Xieping Gao. Msdnn: Multi-scale deep neural network for salient object detection. *arXiv preprint arXiv:1801.04187*, 2018.

[56] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[57] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016.

[58] Lingfeng Zhang, Pengfei Dou, and Ioannis A Kakadiaris. Patch-based face recognition using a hierarchical multi-label matcher. *Image and Vision Computing*, 73:28–39, 2018.

[59] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):611–624, 2019.

[60] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019.

[61] Jian Zhao, Lin Xiong, Panasonic Karlekar Jayashree, Jianshu Li, Fang Zhao, Zhecan Wang, Panasonic Sugiri Pranata, Panasonic Shengmei Shen, Shuicheng Yan, and Jiashi Feng. Dual-agent gans for photorealistic and identity preserving profile face synthesis. In *Advances in Neural Information Processing Systems*, pages 66–76, 2017.

[62] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.