Identity-Preserving Face Anonymization via Adaptively Facial Attributes Obfuscation

Jingzhi Li[1,2], Lutong Han[1,2], Ruoyu Chen[1,2], Hua Zhang[1,2*], Bing Han[1,2], Lili Wang[3], Xiaochun Cao[1,2]
[1] State Key Laboratory of Information Security, Institute of Information Engineering, CAS
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University
Email:{lijingzhi,hanlutong,chenruoyu,zhanghua,hanbing}@iie.ac.cn,wanglily@buaa.edu.cn,caoxiaochun@iie.ac.cn

ABSTRACT

With the popularity of using computer vision technology in monitoring system, there is an increasing societal concern on intruding people's privacy as the captured images/videos may contain identity-related information e.g. people's face. Existing methods on protecting such privacy focus on removing the identity-related information from faces. However, this would weaken the utility of current monitoring system. In this paper, we develop a face anonymization framework that could obfuscate visual appearance while preserving the identity discriminability. The framework is composed of two parts: an identity-aware region discovery module and an identity-aware face confusion module. The former adaptively locates the identity-independent attributes on human faces, and the latter generates the privacy-preserving faces using original faces and discovered facial attributes. To optimize the face generator, we employ a multi-task based loss function, which consists of discriminator loss, identify preserving loss, and reconstruction loss functions. Our model can achieve a balance between recognition utility and appearance anonymizing by modifying different numbers of facial attributes according to pratical demands, and provide a variety of results. Extensive experiments conducted on two public benchmarks Celeb-A and VGG-Face2 demonstrate the effectiveness of our model under distinct face recognition scenarios.

CCS CONCEPTS

• **Computing methodologies → Computer vision**; **Computer vision representations**; **Biometrics**.

KEYWORDS

Privacy preserving, face anonymization, facial obfuscation

ACM Reference Format:
Jingzhi Li[1,2], Lutong Han[1,2], Ruoyu Chen[1,2], Hua Zhang[1,2*], Bing Han[1,2], Lili Wang[3], Xiaochun Cao[1,2]. 2021. Identity-Preserving Face Anonymization via Adaptively Facial Attributes Obfuscation. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475367

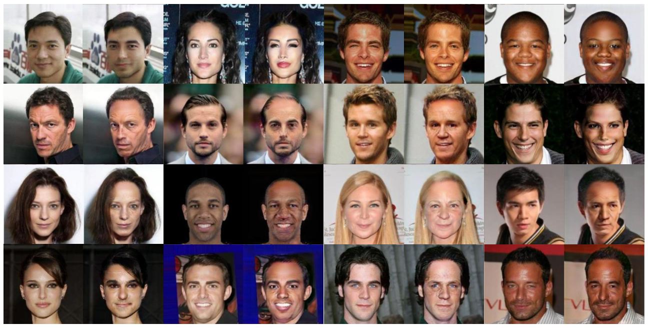


Figure 1: Illustration of our proposed face anonymization method. Our method is able to adaptively discover the identity-independent visual attributes, and then conditioned on these visual attributes the privacy-preserving face is generated. The new face images could be utilized for monitoring system, as their identity related feature representations are almost unaltered. While the observer can not determine weather these two faces referring to the same person, as the facial attributes has been changed.

1 INTRODUCTION

The mass availability of monitoring devices has recoded an amount of facial image data, and many AI-based computer vision technologies are used to mine the personal information at a large scale. Thus, the privacy concerns are growing as the tremendous progress on computer vision technologies. To avoid the abuse of privacy data, some restrictive laws and regulations, e.g. the General Data Protection Regulations (GDPR) [1], require the consent from the individual for any use of their personal data. However, the leakage of facial image data is occurring frequently in world. Moreover, user's facial images stored in the database, even if they are not exposed, are still vulnerable to third-party users or applications. Therefore, face anonymization has become one of the critical steps for many facial applications.

Anonymizing facial images is a challenging task, which requires a robust model to modify the original face without destroying the existing data distribution. Existing methods [2–8] aim to remove all the identification information, and then generate a highly realistic face. These techniques reduce the privacy risks of unnecessary identification, but also destroy the convenience and safety for face

[*] corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475367

recognition. Thus, there is an requirement for retaining the recognition utility in monitoring database, driving license database and other applications.

To satisfy the identity-preserving requirement, Ross et al. [9] introduce the concept of soft biometric privacy, and then develop a face mixing approach to conceal the gender attribute while preserving the recognition utility. In their later studies, adversarial perturbation-based methods [10, 11] are introduced to improve visual quality of generated images which have less artifact. Similarly, Chhabera et al.[12] have adopted the adversarial perturbation technique to realize identity-preserving k-attributes anonymization. Subsequently, PrivicyNet [13] is developed to protect the multi-attributes (gender, race and age). This model could be improve the generalizability of the generated images across multiple face and attribute classifiers. Differently, SensitiveNet [14] is proposed to eliminate sensitive information from the learned embedding space, which prevent privacy leakage for face recognition systems. Although these models have shown their ability on preserving the identity for recognition, the risk on invading the user's privacy is still existing since the original face images are almost unaltered. For practical use, Li et al.[15] have proposed a face camouflage model to preserve the recognition utility under surveillance scenarios. However, this model requires a reference face, which would encourage the reduction of intra-class variations. Thus, the challenge of this technology is how to ensure the face image is only used for authentication without revealing other information about the face.

To tackle the challenge, we develop a novel face anonymization model from the human cognitive perspective, which could obfuscate the visual appearance while retaining the utility for recognition. We first introduce the identity-aware module to discover the facial attributes that closely connected with identity of the individuals. And then, the original face and the discovered facial attribute indicator are fed into the face generator to obtain the privacy-protected face. Specifically, we first compute the identity-aware activation heatmaps of the input face, discovering the regions that support the identity predictions. Next, we parse the input face image into five parts (hair, eyebrows, eyes, nose, and lips), and then associate each part with the corresponding semantic attribute. And then, the identity-independent semantic attributes are obtained via summing the scores of each part and the attribute indicator is obtained. After that, the original face and the attribute indicator are together fed into the conditional face generator for face anonymization. To optimize the proposed model, a multi-task based loss function is employed. It is composed of the discriminator loss, the reconstruction loss, and the identity preserving loss. Extensive experiments conducted on two public benchmarks Celeb-A [16] and VGG-Face 2 [17] demonstrate the effectiveness of our model on anonymizing visual appearance and validate the utility across distinct face recognition scenarios.

Finally, we would reclaim the application scenario of our face anonymization model. In the real scenario, we could deploy the face anonymizer to various applications including surveillance, smart-home cameras, and robots, by designing a lightweight embedded model. As shown in Fig. 1, This would not affect the utility of the face images and protect the individual privacy before uploading the modified images to the database. Furthermore, our model could also generate a cognition gap between the recorded faces and the real faces, which could not be associated by the human observer.

The main contributions of this method are summarized as follows: (1) we develop a face anonymization model to adaptively obfuscate the visual appearance of a face, while preserving the identity discriminability. (2) We present an effective policy for training a deep neural network for privacy protection via a conditional generative adversarial network. (3) We conduct extensive experiments to show the generalization of the proposed approach for multiple face recognizers.

2 RELATED WORKS

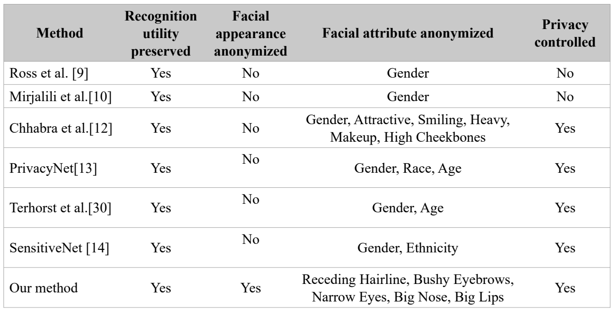
2.1 Face Privacy

Face De-identification Face de-identification [18–20] focuses on preserving facial attributes like gender, age, and race while de-identifying face images, which has evolved over time. Earlier works [21, 22] on face de-identification are mainly naive transformation based. These approaches are the most commonly used in our daily life, and they obfuscate facial sensitive parts through masking, pixelization, blurring and other methods. However, these simple and direct occlusion methods seriously harm the data's availability. What was worse, these methods have been shown to be ineffective with deep learning based face recognition [23]. Another representative methods are the k-same algorithm-based [18]. These algorithms exploit the average face of k-closet faces to replace the given face, which make the face recognition accuracy less than 1/k. Many variants [24, 25] were proposed to improve the data utility and the naturalness of average face. More recently, new techniques and mechanisms have been applied to enhance face privacy. Some researchers implement de-identification by adding adversarial perturbations. Oh et al.[19] introduced a general game theoretical framework to assure user-defined privacy. Shan et al.[26] proposed a system that alter the images' feature space representations using perturbations imperceptible to the naked eye.

GANs inspire a new vein of face de-identification techniques. They can be divided into two categories: those that adopt the conditional inpainting-based technique [20, 27], and those that manipulate facial representations [2, 28]. DeepPrivacy [20] use GAN-based head inpainting technique to generate obscured faces, ensuring privacy-sensitive information is thoroughly removed from the original face. Sun et al.[27] extracts attributes features from the input face, and then generate the anonymous faces. Gafni et al. [2] generate the high-level representations from face images that minimize identity associations, while keeping the perceptions (pose, illumination and expression) unchanged. CIAGAN[7] leverage a vector to control the fake identity of the generated images. IdentityDP [29] combines differential privacy mechanisms with deep neural networks to achieve adjustable privacy control. Specifically, for the identity representation the differential privacy perturbation is added, whereas for the attribute representation is unchanged.

Identity-preserving Face Privacy Identity-preserving face privacy is a newly-developing technology in biometrics and related fields. It tries to prevent other facial information from being used, except the identification purpose. Some researchers have studied the privacy of soft biometric. Ross et al. [9] firstly introduce the

Table 1: The differences between our method and existing identity-preserving face anonymization methods.



Method	Recognition utility preserved	Facial appearance anonymized	Facial attribute anonymized	Privacy controlled
Ross et al. [9]	Yes	No	Gender	No
Mirjalili et al.[10]	Yes	No	Gender	No
Chhabra et al.[12]	Yes	No	Gender, Attractive, Smiling, Heavy, Makeup, High Cheekbones	Yes
PrivacyNet[13]	Yes	No	Gender, Race, Age	Yes
Terhorst et al.[30]	Yes	No	Gender, Age	Yes
SensitiveNet [14]	Yes	No	Gender, Ethnicity	Yes
Our method	Yes	Yes	Receding Hairline, Bushy Eyebrows, Narrow Eyes, Big Nose, Big Lips	Yes

concept of soft biometric privacy, and propose a face mixing approach which conceal the gender attribute while preserving the recognition utility. To improve the results, they have brought forward the adversarial perturbation-based method [10]. Subsequently, Chhabera et al. [12] have also adopted the adversarial perturbation technique to achieve the k-facial attributes anonymization. In [13], Mirjalili et al. develope the Semi-Adversarial Network to impart multi-attribute privacy to face images. This approach can suppress three facial attributes: gender, race and age. Though these soft biometrics privacy methods are shown to successfully prevent the automatic mining of facial attributes, the output images are almost unaltered. In other words, if a human is searching or monitoring the images, they can correctly identify the individual's appearance and other attributes. Recently, some researchers have worked on the privacy of face descriptors. Their goal is to get a facial representation that doesn't contain privacy-sensitive attributes [14, 30]. The facial representation generated by SensitiveNet [14] is eliminated sensitive information such as gender or ethnicity via a modified triplet loss. Terhorst et al.[30] put forward a series of methods to suppress privacy-sensitive information on the template-level. These method enhance the face privacy on representation level instead of the image level for the specific application scenarios. Obviously, storing face representation vectors limited in many applications.

In summary, anonymizing facial appearance is not the main objective of existing identity-preserving face privacy methods. The differences between our method and previous works are shown in Table 1. And Fig. 2 compares the visualization results of different methods. Our method showcases a success in facial appearance anonymization and a significant improvement in confusing human observers.

2.2 Face Manipulation

Face manipulation involves modifying the facial attributes such as the age, the gender, the pose, etc. Among those methods, image-to-image translation algorithms is widely used for real image editing as the flexible adjustment ability. Earlier studies[31, 32] usually perform the image translation in two domains. And then the image translation among multiple domains using only one model is proposed [33, 34]. To improve the attribute manipulation ability, Li et al.[35] propose a novel framework with controllable ability by disentangling the style code in latent space. Similarly, semantic face editing methods can also modify the target facial attributes. Existing works achieved impressive results by designing loss functions[36, 37] and network architectures[38, 39]. InterFaceGAN[40] explore the interpretion of the disentangled face representation, and achieve the controllable manipulation by leveraging the semantic encoded in the latent space. Note that generating visually realistic images is the primary objective of face manipulation methods. Whereas these methods have not considered the recognition utility and privacy protection. These methods cannot be directly used for privacy protection task. Actually, it is challenged to maintain a high recognition performance while obfuscating facial attributes.

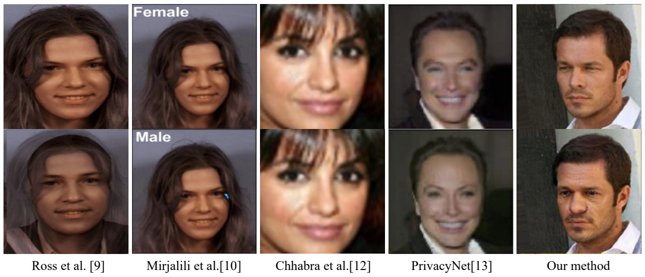


Figure 2: Comparison of our method with existing identity-preserving face anonymization methods. The top row shows original face, and the second row shows the privacy-preserving face.

3 PROPOSED METHOD

Given the original face image $\mathbf{X}_i$, our method aims to generate the corresponding identity-preserving face image $\mathbf{X}_i^p$ in an adaptive manner. The generated face should have the different visual appearance comparing with the original face, but preserve the identity related feature. From the cognition perspective, human is usually sensitive to the semantic variations. Thus, we employ the high perceptual sensitivity facial parts to evaluate the degree of visual appearance variations. Based on related research results [41], the selected facial parts are hair, eyebrows, eyes, nose and lips. For each parts, we select a corresponding facial attribute, which are Receding Hairline, Bushy Eyebrows, Narrow Eyes, Big Nose, Big Lips. The flowchart of our model is presented in Fig. 3. Specifically, we first compute the identity-aware activation heatmaps via the CAM [42], which can be used to localize the identity-related facial parts. After that, we use the face parser [43] to divide the face image into five parts and select the corresponding facial attributes. Then, we compute the response scores for ranking the facial attributes to achieve the identity-independent attribute indicator $\mathbf{L}_t \in \mathbb{R}^c$, where c is the number of attributes. Next, the original face $\mathbf{X}_i$ and the attribute indicator $\mathbf{L}_t$ are fed into the face generator to obtain the final privacy-preserving face image $\mathbf{X}_i^p$.

3.1 Identity-aware region discovery

In this subsection, we aim to determine which facial attributes to be preserved and which to be obfuscated. The preserved attributes represent the critical facial features that closely connected to the identity of individuals. The obfuscated attributes represent the target facial features that loosely connected with the identity of individuals. Actually, the facial attributes of each individual are

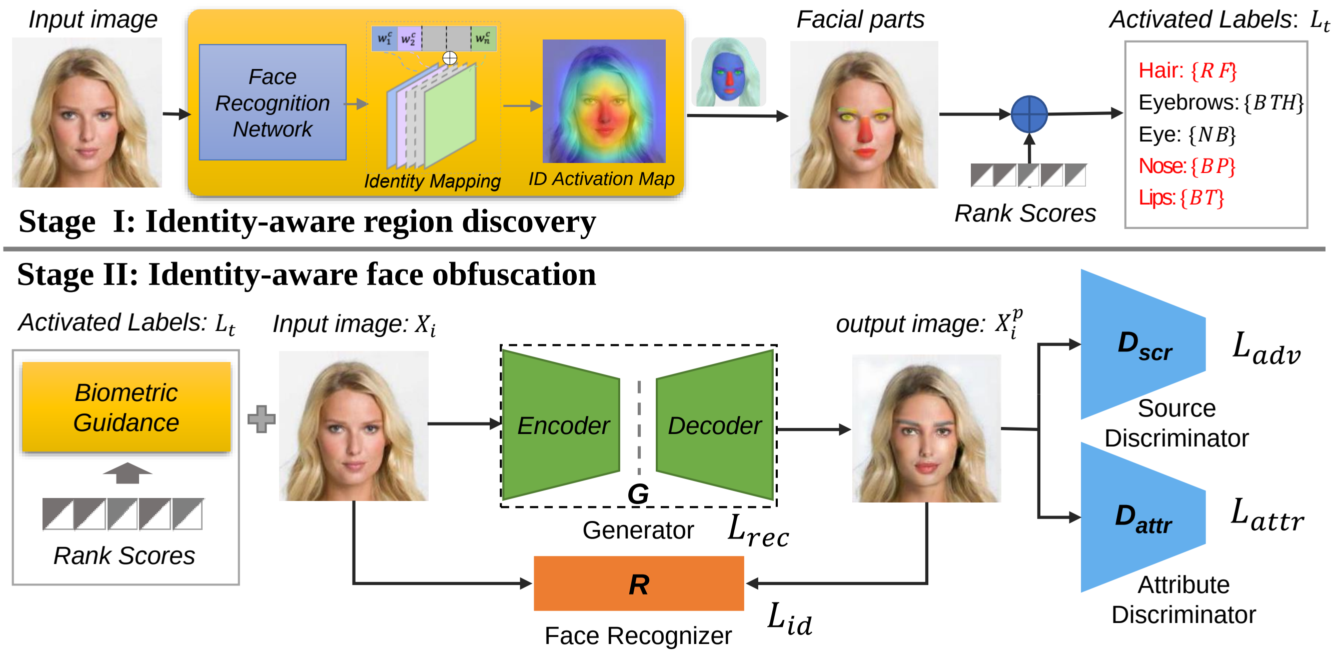


Figure 3: The proposed face anonymization framework. First, the original face image is used to compute the identity-aware class activation heatmap, and then the identity-independent attribute indicator is achieved via ranking the activation scores of each part. After that, the original face and the attribute indicator are fed into the conditional face generator to generate the privacy-preserving face. A multi-task loss function is developed to achieve the end-to-end training procedure, which consists of discriminator loss, the attribute classifier loss, and the face identification loss.

different for the face recognizer and should be learned in some way. Thus, the advantage of this module is that it can adaptively obtain the preserved and obfuscated facial attributes.

Given the presented framework, we first compute the class activation maps, which indicate the discriminative facial parts for face recognition. However, the activation maps can not be directly used for discovering the identity-aware region due to the large receptive field. Therefore, we adopt the face parser to segment the face into facial parts and the face base. The facial parts are the regions containing the eyes, eyebrows, nose, mouth, and hair, which is obtained from the detected facial landmarks. After that, for each facial part, we compute the pixel mean value of the activation maps as the identity relevant scores. Then, we rank the corresponding facial parts in order of scores. Finally, each facial part is labeled with a predefined visual attribute, which is used as the attribute indictor in the face obfuscation module.

The class activation maps of faces are generated by the CAM [42] model. CAM is usually used to obtain a response heatmap for an image classification CNN with a specific type of architecture, where the global average pooled convolutional feature maps are directly fed into the softmax layer. In this work, we first replace the second-to-last fully connected layer of the backbone network with a global average pooling layer and the softmax layer is used for computing the response for the specific identity label. Specifically, the final feature map of face recognition model is denoted as $\mathbf{F} \in \mathbb{R}^{W \times H \times D}$. Where W and H are the width and height of the feature map, and D is the number of feature channels. For an identity class c, the prediction score S_c is obtained from the global average pooling results of the feature maps:

$$S_c = \sum_d \omega_d^c \frac{1}{WH} \sum_{w,h} \mathbf{F}_{d,w,h}, \tag{1}$$

where $\omega \in \mathbb{R}^{C \times D}$ is the weight connecting the feature map. The weight value ω_d^c of the identity class c can be obtained directly from the trained network. Then, the heatmap for identity c is computed as a linear combination of the feature channels:

$$CAM_c(w,h) = \sum_d (\omega_d^c \mathbf{F}_{d,w,h}). \tag{2}$$

By upsampling CAM_c and adding it to the original image, we can get the localization map that highlights the important facial parts. The experimental results are shown in Fig.3. The second and fifth columns in the figure show the identity activation maps, and their next columns are the results added face segmentation. These portraits demonstrate that the identity senstive facial parts are not consistent across different individuals: some individuals' upper facial regions are highly activated, whereas the others are not.

3.2 Identity-aware face obfuscation

In this work, we explore a novel face obfuscation model, which could preserve the underlying identity information and significantly change the visual appearance. Simply considering the image translation is usually not sufficient. We need to find a latent code

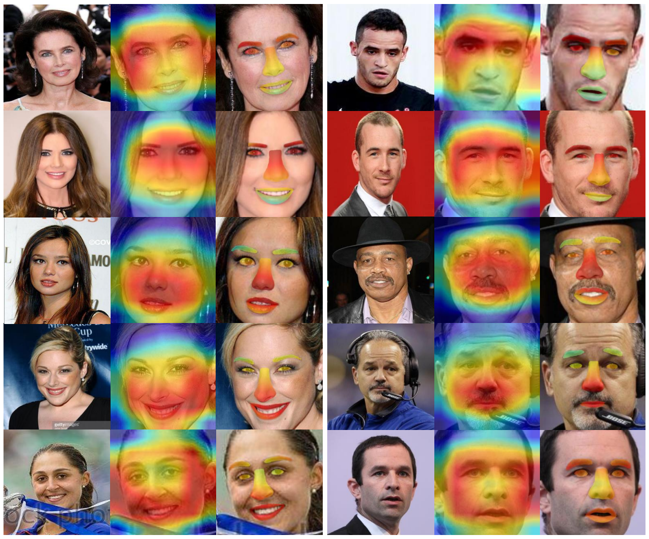


Figure 4: Identity-Sensitive facial parts of different person. The left columns show the original face, the middle columns show the identity activation heatmaps, and the segmentation results of Identity-Sensitive are displayed in the right. For face recognition, the number and locations of Identity-sensitivity facial parts are disparate for different individuals.

that can be used in both appearance variations and face recognition. During this model, the equilibrium between facial features anonymized and recognition utility retained should be achieved. A simple approach to guide face manipulation is through direct latent code optimization based on StarGAN [33], which has been widely used to manipulate the multiple facial attributes.

Fig.3 shows the architecture of this module. It is composed of four sub-networks. The generator **G** is developed to generate the synthesized face with the target label $\mathbf{L_t}$ and the original face. The scource discriminator $\mathbf{D_{src}}$ is trained to distinguish the real images from the synthesized, and the attribute discriminator $\mathbf{D_{attr}}$ is trained to predict the facial attributes of the generated face. To further enhance the recognition utility of generated face, an auxiliary face recognizer **R** is adopted.

Specifically, given the original face X_i and the visual attribute indicator $\mathbf{L_t}$, we solve the following optimization problem:

$$\text{argmin}\, \mathbf{D}_{src}(X_i, \mathbf{G}(\mathbf{X}_i, \mathbf{L}_t) + \mathbf{D}_{attr}(L_t, \mathbf{G}(\mathbf{X}_i, \mathbf{L}_t) + \mathbf{L_{id}}, \quad (3)$$

where **G** denotes the face generator, **D** is to maintain the realistic appearance of synthetic images, and the facial attributes are exactly modified. The term $\mathbf{D}_{src}(.)$ is the probability that the generated image is real or synthesized. The generator **G** and the scource discriminator $\mathbf{D}_{src}$ are trained in the minimax game to iteratively optimize. $\mathbf{D}_{attr}(.)$ returns the probability that the output face image $\mathbf{X_i^P}$ belongs to the target label vector $\mathbf{L}_t$. The generator **G** tries to generate face images that can be properly classified to the target label vector $\mathbf{L_t}$. Moreover, the loss term $\mathbf{L_{id}}$ is used to optimize the performance of the recognition utility preservation.

As described in the previous section, facial attributes consist of preserved attributes and obfuscated attributes. Our model is trained to manipulate the obfuscated attributes and maintain the preserved attributes, which obtained by the visual attribute indicator $\mathbf{L_t}$. The adversarial loss $\mathbf{L_{adv}}$ guides the model to minimize the distribution over real images:

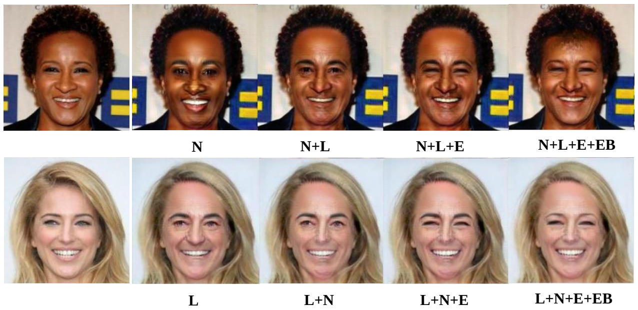


Figure 5: Illustration of the face obfuscation results. Face obfuscation results of person with different numbers of Identity-Sensitive parts. The left columns indicates the original face, and the remaining faces are generated by our model with different attribute indicators. The annotations represent those preserved ID-Sensitive attributes.(EB: Eyebrows, E: Eyes, N: Nose, L: Lips)

$$L_{adv} = E(log(\mathbf{D}_{src}(\mathbf{X_i})) + log(1 - \mathbf{D}_{src}(G(\mathbf{X}_i, \mathbf{L}_t))), \quad (4)$$

The attribute classification loss $\mathbf{L_{attr}}$ is used to optimize the model to achieve the desired attributes:

$$L_{attr} = E(-log(\mathbf{D}_{attr}(L_i|G(\mathbf{X}_i, \mathbf{L}_t)))), \quad (5)$$

where the term $\mathbf{D}_{attr}(L_t|G(\mathbf{X_i}, \mathbf{L_t}))$ indicates a probability of the synthesized face belonging to the class L_t. By minimizing this objective, G learns to generate images that can be classified as desired attributes. The attribute discriminator $\mathbf{D}_{attr}$ was pre-trained by the labeled training data, which is a supervised training process.

To preserve the content with the unselected attributes of the original face image, we minimize the L1 norm as the reconstruction loss:

$$L_{rec} = E(||\mathbf{X_i} - G(G(\mathbf{X}_i, \mathbf{L}_t), \mathbf{L}_o)||_1), \quad (6)$$

where $\mathbf{L}_o$ represents the original attribute label. In other words, **G** was secondly used to reconstruct the original face image from the generated face image in the training stage.

Finally, the recognition consistency of the input face and the generated face is controlled by the identity loss:

$$L_{id} = \sum_{k \in n} ||\mathbf{R_k}(\mathbf{X_i}) - \mathbf{R_k}(\mathbf{X_i^P})||_2, \quad (7)$$

where $\mathbf{L_{id}}$ is the perceptual loss for learning the facial details, **n** is a collection of convolution layers from the perceptual network and $\mathbf{X_i}$ is the activation from the k-th layer. In our work, the perceptual network is pretrained on a face dataset to enforce that the generated face contains recognition features.

The total objective function for our model defined as:

$$L_{total} = L_{adv} + \alpha_{attr} L_{attr} + \alpha_{rec} L_{rec} + \alpha_{id} L_{id}, \quad (8)$$

where α is a trade-off parameter. The first three terms in above equation forces the output image to leave the original image, while the last term encourages the face recognizer **R** to judge the two face images to be the same person. In our method, $\alpha_{\mathbf{id}}$ could vary for different iterations. At first, we set small values to make the generator really change the face image. Then, $\alpha_{\mathbf{id}}$ increases to pull the image back, and the result of **R** from the generated face is as the same as that from the original face.

In Fig.5, we show the face sequences generated by the face obfuscation model. Starting from the second face of each row, the number of facial attributes modified decreases. We can observe the facial appearance have significantly changed.

4 EXPERIMENTS

We adopt two public benchmarks Celeb-A [16] and VGG-Face 2 [17] to evaluate the effectiveness of our model. Three types of experiments are performed. For the facial appearance anonymization, we analyze the performance on the diversity and realism of generated images. For the identity preservation, we analyze the performance of generated images on three face recognizers. In addition, a user study is added to verify the effectiveness of this method for human observers.

4.1 Datasets

Celeb-A [16] is one of the most widely used datasets for face attribute recognition. It contains 202, 599 images of 10, 177 identities. Each image has 5 landmark locations and 40 attribute annotations. The original face images are first cropped into 178 × 178, and then resized to 128 × 128. Among them, we randomly select 2,000 images as the test set and the remaining images as the training set.

VGG-Face 2 [17] contains 3.31 million images of 9, 131 subjects, with an average of 362.6 images for each subject. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession (e.g. actors, athletes, politicians). In particular, this dataset is divided into the training and test set, which a training set consists of 8, 631 identities (3, 141, 890 images) and a test set contains 500 identities. We randomly select one image for each identity from the test set.

4.2 Implementation details

For data processing, we follow recent models [44, 45] to crop and align faces to get the normalized face images (112 × 112). For face obfuscation model, we employ StarGAN [33] as the basic generation network. In terms of network architecture, we adopt a generate network that contains two convolutional layers for downsampling, six residual blocks, and two transposed convolutional layers for upsampling. The model is trained using Adam with $\beta 1 = 0.5$ and $\beta 2 = 0.999$, and we perform one generator update after five discriminator updates. The batch size is set to 16 for all experiments. In the experiment on CelebA, we train the model with a learning rate of 0.0001 for the first 10 epochs, whereas the learning rate linearly decay to 0 over the next 10 epochs. For identity preservation task, we perform the black box attack on three face recognizers. Specifically, we use the recent state-of-the-art model (Cosface [44], Arcface [46] and VGG-Face 2 [17]) to evaluate the performance of face recognition. Furthermore, we explore the batch normalization and dropout layer to train the robust face recognizer.

Training Strategy. When training the face generator with multiple attributes, we use the facial attribute indicator and the original face as input. In the meantime, the generator should learn to ignore the unspecified visual attributes, which are set to zero vectors, and focus on the explicitly specific attributes. The structure of the generator is exactly the same as StarGAN, expect for the dimension of the input attribute indicator. On the other hand, we extend the auxiliary classifier of the discriminator to predict the probability distributions over attributes for all datasets. Then, we train the model in a multi-task learning setting, where the discriminator tries to learn the discriminative features of both datasets, and minimize the errors associated with the attributes. Under these settings, the generator learns to control all the visual attributes of both datasets.

4.3 Evaluation metrics

LPIPS Distance [47] is used to measure the similarity between the generated images and the original images. We employ the average LPIPS distance between pairs of randomly-sampled outputs from the same input. LPIPS is computed based on a weighted L_2 distance between deep features of images. It has been demonstrated to correlate well with human perceptual similarity [47]. Following [48], we use all the test images and the corresponding output pairs for each input, and then extract their feature representation based on the pre-trained face recognition model.

FID Distance [49] is used to calculate the distance between the distribution of the real images and the generated images. The lower FID score denotes that the quality of generated images is higher and more diverse to the real ones. Thus, we follow the experimental setting as [49] to evaluate the performance on the realism of generated images.

Human Preference. To compare the realism and faithfulness of generated images by our model, we introduce human perceptual study. Similar to [50], we present on input image and the generated images by our model to the workers. They are given unlimited time to select which pairs are significant different. For each comparison, we randomly select different workers to answer it and use the average score as the ground-truth.

4.4 Evaluation on appearance anonymization

The visual appearance anonymization of our model is evaluated on both Celeb-A and VGG Face 2 datasets. Modifying different numbers of facial attributes is considered in this task. We show the quantitative and qualitative experimental results in Table 2 and Fig. 6. For quantitative evaluation, we adopt LPIPS and FID to evaluate the visual varations and realism of generated images. The former is used to assess the perceptual similarity between the original face and the generated face. The latter is sensitive to visual artifacts, which can indicate the realism of generated images. Our goal is to anonymize the facial performance, in terms of visual quality and diversity, we only need a practical level. Hence, we only employ the StarGAN [33] as the comparison baseline. From the experimental results in Table 2, we observe that our model achieves a better performance under the LPIPS and FID metric on both two benchmarks. Meatime, we could find that the LPIPS and FID values get

Table 2: Quantitative comparisons on realism and diversity of synthetic images between our model and StarGAN on CelebA and VGG Face 2 datasets.(1 ID-Sensitive part indicates that one facial attribute has been obfuscated.)

Method	1 ID-Sensitive part		2 ID-Sensitive parts		3 ID-Sensitive parts		4 ID-Sensitive parts		adaptive	
	LPIPS ↑	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑	FID ↓
StarGAN[33]	0.108	82.146	0.130	73.077	0.134	63.092	0.118	52.038	0.122	66.112
our model (VGG-Face)	0.112	78.307	0.135	78.307	0.137	60.189	0.121	48.808	0.134	64.512
StarGAN[33]	0.086	9.313	0.103	14.15	0.103	20.07	0.087	26.50	0.087	26.502
our model (Celeb-A)	0.088	8.835	0.104	12.41	0.104	12.41	0.087	22.84	0.114	22.132

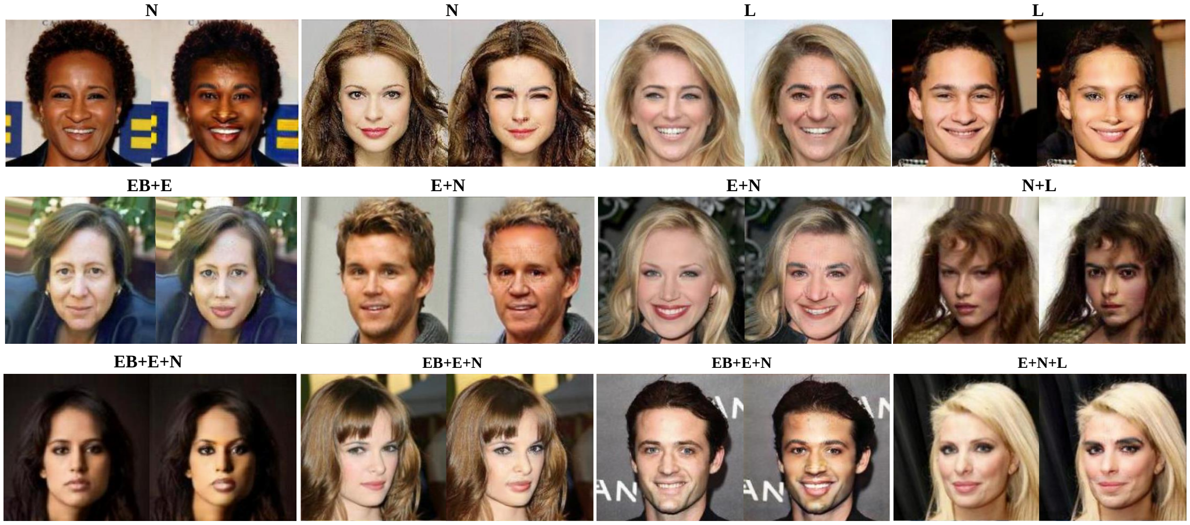


Figure 6: Privacy protection results for individuals with different numbers of ID-Sensitive parts. All images are sampled from the test set. The left columns show the un-protected face, the right columns show the protected face generated by our method. Despite the significant changes in appearance, each pair was still identified as the same person. The annotations represent the preserved ID-Sensitive attributes.(EB: Eyebrows, E: Eyes, N: Nose, L: Lips)

better with the number of attributes modified. Furthermore, when using the adaptive mechanism to generate the privacy-preserving face image, our model still achieves expected results, and does not reduce the realism and diversity of the generated images.

Fig. 6 show the visual experimental results with different numbers of facial attributes modified. This demonstrates that our model has the ability to generate the realistic images, which have a significant visual variation comparing with the original face. Some typical experimental results are shown in Fig. 5. For each pair, the first column is the original and other columns are generated by our framework. The results show that our proposed model can generate new faces with different visual appearances but identified as the same identity.

4.5 Evaluation on identity preservation

To evaluate the performance on identity-preserving while anonymozing visual appearance, we conduct some experiments on both Celeb-A and VGG Face 2 datasets. Experiments of the black box attack have been performed on three face recognizers. Three widely used face recognizers: ArcFace [46], CosFace [44] and VGG-Face 2 [17], are used to match the original to anonymized faces. Specifically, on Celeb-A dataset we randomly selected 1, 000 same identities' face pairs and 2, 000 different identities' face pairs to compute the average accuracy of face verification by the face recognizers. While on VGG-Face 2 dataset, we randomly selected 500 same identities' face pairs and 1, 000 different identities' face pairs to develop the testing list. The experimental results are shown in Table 3 and 4. Our model can generate 4 images for each input face according to the number of visual attributes. Meanwhile, the model can automatically find the optimal face image according to the predefinedthreshold value. Hence, we compute the True Accept Rate (TAR) values at two False Accept Rate (FAR) value for these 5 cases. The experimental results show that our model achieves satisfactory average face verification rate on three face recognizers, which show the generalizability under distinct face recognition scenarios.

4.6 User study

To evaluate the validity of our results on confusing human observers, we conduct a user study with two experiments. The first experiment aims to determine the identity related facial parts. To

Table 3: Performance of the identity-preserving face anonymization with different number of visual attributes and distinct face recognizers on VGG Face 2 in terms of TAR (%) at FAR = 0.1 and 0.01.(1 ID-Sensitive part indicates that one facial attribute has been obfuscated.)

Method	1 ID-Sensitive part		2 ID-Sensitive parts		3 ID-Sensitive parts		4 ID-Sensitive parts		adaptive	
	0.01	0.1	0.01	0.1	0.01	0.1	0.01	0.1	0.01	0.1
ArcFace [46]	0.6751	0.8795	0.4568	0.7405	0.3449	0.6384	0.2703	0.5622	0.4794	0.7312
VggFace [17]	0.7178	0.8811	0.5211	0.7500	0.3468	0.6157	0.2789	0.5232	0.5186	0.7442
CosFace [44]	0.8497	0.9508	0.6695	0.8762	0.5300	0.7862	0.4578	0.7292	0.6128	0.7713

Table 4: Performance of the identity-preserving face anonymization with different number of visual attributes and distinct face recognizers on Celeb-A in terms of TAR (%) at FAR = 0.1 and 0.01.(1 ID-Sensitive part indicates that one facial attribute has been obfuscated.)

Method	1 ID-Sensitive part		2 ID-Sensitive parts		3 ID-Sensitive parts		4 ID-Sensitive parts		adaptive	
	0.01	0.1	0.01	0.1	0.01	0.1	0.01	0.1	0.01	0.1
ArcFace [46]	0.8304	0.9689	0.6505	0.9051	0.5015	0.8250	0.3714	0.7442	0.6439	0.7523
VGG Face [17]	0.9366	0.9870	0.8133	0.9577	0.7718	0.9497	0.6636	0.9034	0.7312	0.8410
CosFace [44]	0.9601	0.9936	0.8830	0.9797	0.5049	0.8273	0.6706	0.9053	0.7839	0.8561

Table 5: Experimental results of user study on human perceptual facial attribute and face verification.

Human-perceptual Facial Attributes					Verification	
hairline	eyebows	eyes	nose	lips	Real	Fake
0.12	0.34	0.67	0.38	0.59	0.98	0.88

that end, we randomly select 200 face images from different identities, and set five options (hairline, eyebrows, eyes, nose, and lips) for each face. Users are requested to select the facial parts out of our options, they can make single choice or multiple choices (no more than 4 choices). We collected a total of 5000 results from 25 volunteers, and record the number of different facial parts. The probability of each facial attribute is reported in Table 5. The results indicate that two facial parts (eyes and lips) are selected at high frequency, and the other three at low frequency. From the human perspective, changing high-frequency facial parts will make faces look like different people, while the low-frequency attributes have weak correlation with identity, especially the hairline. Therefore, we did not consider to modify the hairline in the synthesis process.

The second experiment aims to validate if the generated face would confuse the users. Specifically, we randomly select 200 image pairs composed of 100 males and 100 females, with an equal number of pairs those belonged to the same person and those not. Then, we collected a total of 5000 results from 25 volunteers. Finally, we calculate the average accuracy of face verification to measure the ability of visual appearance anonymization, as shown in Table 5. The results show that the anonymous face generated by our method could significantly reduce the recognition rate of human.

5 CONCLUSION

In this paper, we develop a novel framework for protecting the privacy of face images in monitoring system. We introduce a novel face anonymization model, which combine the strong generative powers of StarGAN with the discovered extraordinary facial attribute indicators. We have shown that our model has the ability to generate a wide appearance variations of face images with the identity-preserving. Experimental results demonstrate that our method can preserve the recognition utility for distinct face recognizers, and effectively anonymize the facial appearance. We have also demonstrated that our method provides fine-grained edit controls, such as specifying a desired attribute e.g. big nose. In the future, we will explore our model's ability to protect other facial attributes, eg. race, age, emotion, etc.

ACKNOWLEDGEMENT

Supported by the National Key R&D Program of China (Grant No. 2018AAA0100601), National Natural Science Foundation of China (No. 62025604, 62072454, U1936210), Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (No.VRLAB2021C06), Beijing Natural Science Foundation (No.4202084).

REFERENCES

[1] Razvan Viorescu et al. 2018 reform of eu data protection rules. *European Journal of Law and Public Administration*, 4(2):27–39, 2017.

[2] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9378–9387, 2019.

[3] Hanxiang Hao, David Güera, Amy R Reibman, and Edward J Delp. A utility-preserving gan for face obscuration. *arXiv preprint arXiv:1906.11979*, 2019.

[4] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[5] Xiuye Gu, Weixin Luo, Michael S Ryoo, and Yong Jae Lee. Password-conditioned anonymization and deanonymization with face identity transformers. In *European Conference on Computer Vision*, pages 727–743, 2020.

[6] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[7] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5447–5456, 2020.

[8] Shuhui Yang, Han Xue, Jun Ling, Li Song, and Rong Xie. Deep face swapping via cross-identity adversarial training. In *International Conference on Multimedia Modeling*, pages 74–86, 2021.

[9] Asem Othman and Arun Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *European Conference on Computer Vision*, pages 682–696. Springer, 2014.

[10] Vahid Mirjalili and Arun Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *2017 IEEE International joint conference on biometrics (IJCB)*, pages 564–573. IEEE, 2017.

[11] Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)*, pages 82–89. IEEE, 2018.

[12] Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. Anonymizing k-facial attributes via adversarial perturbations. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 656–662, 2018.

[13] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Privacynet: semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412, 2020.

[14] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[15] Jingzhi Li, Lutong Han, Hua Zhang, Xiaoguang Han, Jingguo Ge, and Xiaochun Cao. Learning disentangled representations for identity preserving surveillance face camouflage. In *25th International Conference on Pattern Recognition*, 2020.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[17] 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, 2018.

[18] Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[19] Seong Joon Oh, Mario Fritz, and Bernt Schiele. Adversarial image perturbation for privacy protection a game theory perspective. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1491–1500. IEEE, 2017.

[20] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International Symposium on Visual Computing*, pages 565–578. Springer, 2019.

[21] Carman Gerard Neustaedter and Saul Greenberg. *Balancing privacy and awareness in home media spaces*. Citeseer, 2003.

[22] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 1–10, 2000.

[23] Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *2006 Conference on computer vision and pattern recognition workshop (CVPRW'06)*, pages 161–161. IEEE, 2006.

[24] Liang Du, Meng Yi, Erik Blasch, and Haibin Ling. Garp-face: Balancing privacy protection and utility preservation in face de-identification. In *IEEE International Joint Conference on Biometrics*, pages 1–8. IEEE, 2014.

[25] Amin Jourabloo, Xi Yin, and Xiaoming Liu. Attribute preserved face de-identification. In *2015 International conference on biometrics (ICB)*, pages 278–285. IEEE, 2015.

[26] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1589–1604, 2020.

[27] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5050–5059, 2018.

[28] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. Infoscrub: Towards attribute privacy by targeted obfuscation. *arXiv preprint arXiv:2005.10329*, 2020.

[29] Yunqian Wen, Li Song, Bo Liu, Ming Ding, and Rong Xie. Identitydp: Differential private identification protection for face images. *arXiv preprint arXiv:2103.01745*, 2021.

[30] Philipp Terhörst, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Unsupervised privacy-enhancement of face representations using similarity-sensitive noise transformations. *Applied Intelligence*, 49(8):3043–3060, 2019.

[31] 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.

[32] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[33] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[34] 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*, 28(11):5464–5478, 2019.

[35] Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. *arXiv preprint arXiv:2103.01456*, 2021.

[36] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1415–1424, 2017.

[37] Yaxing Wang, Abel Gonzalez-Garcia, Joost van de Weijer, and Luis Herranz. Sdit: Scalable and diverse cross-domain image translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1267–1276, 2019.

[38] Yujun Shen, Ping Luo, Junjie Yan, Xiaogang Wang, and Xiaoou Tang. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 821–830, 2018.

[39] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[40] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[41] Galit Yovel Naphtali Abudarham, Lior Shkiller. Critical features for face recognition. *Cognition*, 182, 2019.

[42] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[43] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Hierarchical face parsing via deep learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2480–2487. IEEE, 2012.

[44] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[45] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[46] Jiankang Deng, Jia Guo, Xue Niannan, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019.

[47] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[48] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *arXiv preprint arXiv:1711.11586*, 2017.

[49] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:1706.08500*, 2017.

[50] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.