Occluded Face Recognition in the Wild by Identity-Diversity Inpainting

Shiming Ge, *Senior Member, IEEE,* Chenyu Li, Shengwei Zhao, and Dan Zeng

***Abstract*—Face recognition has achieved advanced development by using convolutional neural network (CNN) based recognizers. Existing recognizers typically demonstrate powerful capacity in recognizing un-occluded faces, but often suffer from accuracy degradation when directly identifying occluded faces. This is mainly due to insufficient visual and identity cues caused by occlusions. On the other hand, generative adversarial network (GAN) is particularly suitable when it needs to reconstruct visually plausible occlusions by face inpainting. Motivated by these observations, this paper proposes identity-diversity inpainting to facilitate occluded face recognition. The core idea is integrating GAN with an optimized pre-trained CNN recognizer which serves as the third player to compete with the generator by distinguishing diversity within the same identity class. To this end, a collect of identity-centered features is applied in the recognizer as supervision to enable the inpainted faces clustering towards their identity centers. In this way, our approach can benefit from GAN for reconstruction and CNN for representation, and simultaneously addresses two challenging tasks, face inpainting and face recognition. Experimental results compared with 4 state-of-the-arts prove the efficacy of the proposed approach.**

***Index Terms*—Occluded face recognition, Inpainting, GAN, Deep learning.**

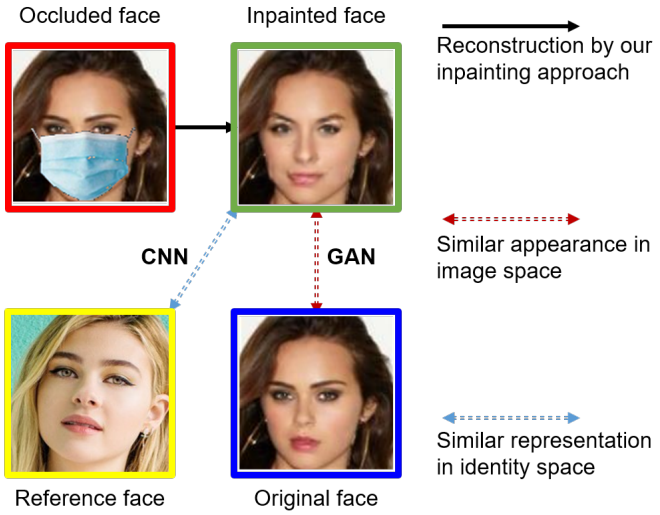

Fig. 1. Motivation of our approach. We combine the capacity of GAN-based reconstruction and CNN-based recognition to generate realistic and identity-preserving faces.

I. Introduction

OCCLUSION is a common issue when recognizing faces in real-world applications like video surveillance [1]. In spite of impressive performance achieved by many well-known deep recognizers [2], [3], [4], [5], [6], [7] on un-occluded faces, the accuracy may have a sharp drop when directly recognizing occluded ones. Different from un-occluded faces, occluded faces are difficult to be recognized due to incomplete visual content and insufficient identity cues. However, they are still recognizable for subjects when giving partial facial content, implying that the neural systems of human beings may have the capability of recovering missing facial cues from occlusions. Inspired by this fact, many existing occluded face recognition approaches have been proposed, which mainly base on two ideas: representation and reconstruction.

The "representation" approaches aim to directly represent the occluded faces with features extracted from partial facial content. Wright [8] proposed extracting facial features robust to occlusion by sparse representation. Yang *et al.* [9] extended this idea and proposed representing faces by modeling a sparsity-constrained regression problem. Similar approaches were also proposed by adding low-rank regularization [10] and extracting dynamic subspace representation [11]. Wei *et al.* [12] proposed to separate the face region into several ordered parts and represented it by dynamic image-to-class warping. By using the powerful representation ability of convolutional neural network (CNN), the approach in [13] treated difference due to various poses and occlusions as multi-modality of faces, and proposed a conditional CNN to handle multi-modal face representation. Recent deep face recognizers employed various loss functions for training, *e.g.*, triplet loss [4] and center loss [5], to improve the robustness against occlusions. These approaches achieved good performance in

Manuscript received xx xx, 20xx; accepted xx xx, 20xx. Date of publication xx xx, 20xx; date of current version xx xx, 20xx. This work was partially supported by grants from the National Natural Science Foundation of China (61772513). Shiming Ge is also supported by the Open Projects Program of National Laboratory of Pattern Recognition, and the Youth Innovation Promotion Association, Chinese Academy of Sciences.(*Corresponding authors: Shiming Ge and Dan Zeng*)

S. Ge is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China. E-mail: geshiming@iie.ac.cn.

C. Li and S. Zhao are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China, and School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China. Email: {lichengyu, zhaoshengwei}@iie.ac.cn.

D. Zeng is with Key laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute of Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China. E-mail: dzeng@shu.edu.cn.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier xx.xxxx/TCSVT.xxxx.xxxxxx

recognizing faces with small occlusions. However, they still suffer from low recognition accuracy under heavy occlusions due to the difficulty in recovering identity cues.

Unlike the "representation" approaches, the "reconstruction" approaches propose to reconstruct the missing facial parts before recognition. Early work proposed by Deng [14] applied graph-based algorithm for repairing face image to improve the recognition performance on occluded faces, where sparse representation was used for feature extraction. Traditional exemplar-based approaches [15], [16] applied texture synthesis to fill in the missing parts with the nearest available exemplars. Recently, with the help of generative adversarial network (GAN) [17], face inpainting approaches [18], [19], [20], [21], [22] have achieved remarkable improvement in extracting high-level context features and generating photorealistic results. Beyond realism improvement in the generated faces, some GAN models [23], [7] took identity information into consideration to perform identity-preserving inpainting. The flexility of the framework as well as the realism of the generated faces make GAN a feasible fit for occluded face recognition, where effective transferring of identity knowledge plays a important role.

In this work, we address the recognition of occluded faces by identity-diversity inpainting, which integrates a CNN-based face recognizer into GAN, and propose Identity-Diversity GAN (ID-GAN). Therefore, ID-GAN can benefits from both sides, GAN for visual reconstruction and CNN for feature reconstruction, which allows its powerful capacity in generating photorealistic and identity-preserving results. ID-GAN treats the superviser as the third player to interact with the other two players, competing with the generator by demanding diversity among faces with the same identity label, while cooperating with the discriminator by punishing discriminativeness between inpainted and ground-true faces. Toward this end, the superviser employs a collect of identity-centered features as supervision signals to regularize the generator by maximizing the aggregation of identity features. In this way, the generator is forced to reduce the difference with ground-truth in both pixel and feature spaces and improve the recognition capacity by enhancing identity diversity. Our main contributions can be summarized as three folds:1) we propose identity-diversity inpainting by integrating GAN with a well-trained face recognizer, which servers as the third player and facilitates occluded face recognition when giving photorealistic results, 2) We propose an identity-diversity loss with the supervision of a collect of identity-centered features, which could suppress identity diffusion and improve the discriminative capacity of the generated faces, and 3) We conduct qualitative and quantitative experiments to show the efficacy of the proposed approach on generating photorealistic results and improving occluded face recognition.

II. Related Works

The approach we proposed in this paper aims to address occluded face recognition via inpainting. Therefore, we briefly review related works from four aspects, including deep face recognition models, occluded face recognition, face inpainting and identity-preserving face synthesis.

A. Deep Face Recognition Models

Face recognition is one of the fundamental and also most successful topic in pattern recognition, with long history of research. Traditional face recognition heavily relied on hand-crafted features [24], [25], [26] or fixed dictionaries [27]. Recently, two factors, including end-to-end learning for the task using deep learning and the availability of massive training datasets, have enabled the great process in face recognition areas. Compared to the hand-crafted features, learning-based methods were able to exploit information with better discriminative ability for recognition [28], and therefore shifted the research to the sceneries in the wild. In this literature, several remarkable deep face recognition models [6], [4], [3], [2], [5], [29], [30], [31], [32], [33] have achieved very impressive recognition performance. These models have considered a lot about the network structure and loss function adopted. For example, Xiong *et al.* [34] adopted deep mixture model and convolutional funsion network(DMM+CFN). [35] proposed to improve verification performance by measuring similarity between two image sets instead of two individual images. As for loss functions, DeepFace [2] is an early attempt to ensemble CNN by building 3D faces with identification loss. After that, various loss functions have been proposed for training face recognition models, such as triplet loss [6], [4], [36], center loss [5], Sphere loss [30], range loss [31], Cos Loss [32] and ring loss [33]. In [37], Zhu *et al.* proposed large-scale bisample learning (LBL) method to address ID versus Spot face recognition problem that is challenged by the intra-class variations in two face samples for each class.

In general, these face recognition models mainly are trained on massive un-occluded faces, and thus will achieve good accuracy when applied on recognizing the faces after inpainting.

B. Occluded Face Recognition

Existing occluded face recognition approaches mainly are grouped into two categories: representation-based and reconstruction-based. The representation-based approaches aim to implicitly represent faces by devising a feature extractor robust to occlusion, while the reconstruction-based approaches explicitly recover the occluded facial parts and then apply the reconstructed faces for recognition.

The representation-based approaches represent occluded faces by extracting features robust to occlusion in a local or global way. In the local way, some approaches first segmented a face image into several local parts and then described the face by using the ordered property of facial parts [5] or extracting discriminative components [11]. Unlike the local way, global way directly takes the whole face image as input and represent it with a good descriptor, such as sparse representation [9] and low-rank regularization [10]. Motivated by the powerful ability of CNN, Xiong [13] proposed a conditional CNN to learn face representation by dynamically activated sets of convolutional kernels. In addition, many CNN-based face recognition models [2], [3], [4], [5], [6] showed robust representation by learning from massive data. In general, these approaches focus on reducing the effect of occlusion in representing faces

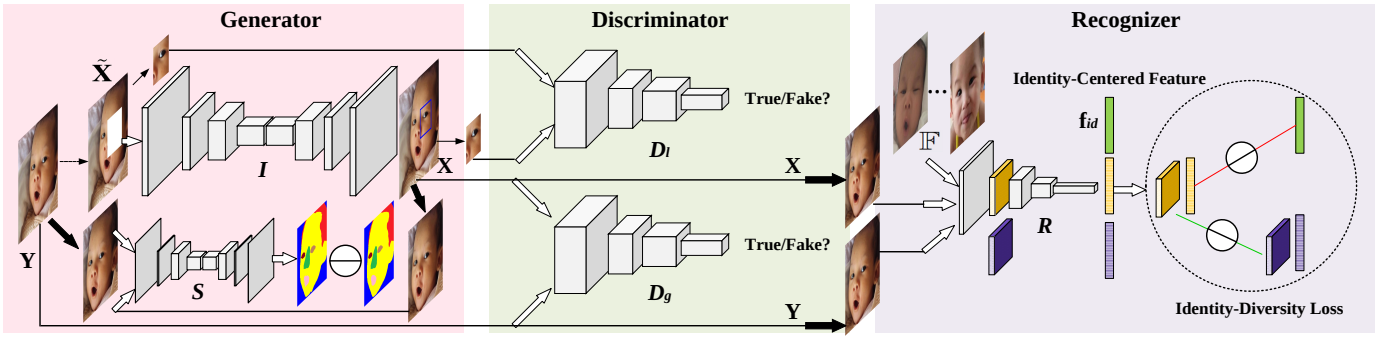

Fig. 2. The ID-GAN framework. It aims to produce photorealistic and identity-preserving faces via the cooperation and competition among three parties: 1) the generator uses an inpainting network $\boldsymbol{I}$ to synthesis the inpainted face $\mathbf{X}$ while preserves structure fidelity by a structure network $\boldsymbol{S}$, 2) the discriminator jointly identifies the real or fake of $\mathbf{X}$ by a local discriminative network $\boldsymbol{D}_l$ and a global discriminative network $\boldsymbol{D}_g$, and 3) the recognizer recovers identity cues via a recognition network $\boldsymbol{R}$ which constructs an identity-diversity loss involving an identity triplet: the inpainted face $\mathbf{X}$, the ground-truth face $\mathbf{Y}$ and an identity-centered feature $\mathbf{f}_{id}$. In this way, the rich identity knowledge can be transferred from $\boldsymbol{R}$ to $\boldsymbol{I}$. Finally, $\boldsymbol{I}$ and $\boldsymbol{R}$ are cascaded to recognize the occluded face $\tilde{\mathbf{X}}$. Thus, the framework actually addresses two challenging tasks, occluded face inpainting and occluded face recognition, in a unified framework. Note that the face images are cropped and resized (marked in big solid arrow) to adapt $\boldsymbol{S}$ and $\boldsymbol{R}$.

without extra knowledge introduced, leading to insufficient feature representation when having large occlusion.

The reconstruction-based approaches tend to construct the missing facial parts before recognition. Deng *et al.* [14] proposed a graph Laplace algorithm to recover occluded faces, which boosts recognition accuracy. More recent GAN-based approaches were proposed to synthesis occluded faces. Zhang *et al.* [23] proposed DeMeshNet to enforce pixel as well as feature-level similarity between input and output face images. It can recover the missing contents with little pixel difference and improve recognition performance. To deal with large occluded region, Li *et al.* [20] proposed GAN to perform face inpainting by introducing segmentation prior, which improves the accuracy of face verification. Zhao *et al.* [7] further proposed a dual-agent GAN model to improve realism of face images, while preserving identity information during realism refinement by introducing an identity perception loss into GAN. Typically, the reconstruction-based approaches introduce new knowledge during the reconstruction and give appealing performance. Now, a key question is that how to make the introduced knowledge suitable for face recognition.

C. Face Inpainting

Face inpainting aims to fill-in the missing facial parts. By using the available parts as reference, early exemplar-based approaches [15], [16], [38] searched similar patches to synthesis the missing regions. This non-parametric manner could achieve good results when having available content but lead to unnatural outputs on inpainting faces with unique textures. An alternative is using parametric way [39], which learns dictionary to recover large area. Generally, these approaches are difficult to generate semantic content due to low-level processing manner and lack of introduced knowledge.

Recent GAN-based approaches could well synthesis semantic content by learning from massive data. Pathak *et al.* [18] proposed context encoder to learn feature representation to capture both appearance and visual semantics. Iizuka *et al.* [19] employed two discriminators to jointly enforce global and local consistency, leading to a visually natural result. Wang *et al.* [40] employed perceptual loss to increase semantic similarity. In [22], dilated convolution layers were used in GAN to improve inpainting efficiency. Zhao [41] proposed restoring partially occluded faces in the wild via a robust LSTM-Autoencoders model where two LSTM components were used to occlusion-robust face encoding and recurrent occlusion removal respectively. In summary, GAN framework gives an opportunity to introduce high-level semantic information (*e.g.*, identity) besides visual improvement.

D. Identity-Preserving Face Synthesis

Identity-preserving face synthesis techniques aim to generate highly realistic faces while maintaining the identity information. They have been widely applied in varieties of computer vision tasks for improving face recognition.

In [42], Berg *et al.*proposed to take an extra reference face set to perform identity-preserving alignment by warping the faces to reduce differences due to pose and expression, thereby improving face verification performance. Zhu *et al.* [43] proposed learning identity-preserving features in the canonical view to address face recognition challenges. They designed a deep network to combine the feature extraction layers and the reconstruction layer. Zhao *et al.* [7] proposed Dual-Agent GAN (DA-GAN) model to improve the realism of a face simulator's output using unlabeled real faces, while preserving the identity information during the realism refinement. Shiri *et al.* [44] proposed face recovery from portraits by combining a style removal network and a discriminative network with the GAN framework, where the identity preservation is ensured by promoting the recovered and ground-truth faces to share similar visual features extracted by a pre-trained VGG network. Dolhansky and Ferrer [45] proposed to inpaint eye for preserving identity via Exemplar GANs (ExGANs), which uses a reference region image or a object perceptual code as exemplar informaiton. Shen *et al.* [46] proposed FaceID-GAN, which was formulated as a third-party competition game by introducing an identity consistency on synthesized

face. Bao *et al.* [47] proposed to recombine the identity vector of an input face and the attribute vector of the other face to achieve identity-preserving synthesis, along with an identity perception loss for distinguishing identities. Wang *et al.* [48] proposed identity-preserving sketch synthesis method by replacing traditional k-NN with anchored neighborhood index of neighbors in terms of distance between both photo patches and sketch patches. In [49], Li *et al.*proposed a bi-level adversarial network to perform de-makeup, where two adversarial networks were integrated in an end-to-end fashion, with one on pixel level for reconstructing facial images and the other on feature level for preserving identity information. Some recent works [50], [51] try to synthesize faces by identity-preserving attribute manipulation, showing impressive performance in cross-age face verification task.

These approaches mainly focus on preserving identity information during synthesis by defining an identity loss. The loss usually plays a role as regularization instead of competition.

III. The Proposed Approach

Our identity-diversity inpainting framework takes an alternative solution to recognize occluded faces by applying an inpainting model to fit the accuracy of "normal" face recognition models. Thus, the inpainting model should improve occluded face recognition by meeting two rules: photorealistic and identity-preserving. Inspired by that, we introduce a well-trained face recognizer into GAN, resulting in Identity-Diversity GAN (ID-GAN) which involves three players (as shown in Fig. 2), including the generator, the discriminators and the recognizer. In this way, the rich identity knowledge contained in the recognizer can be utilized to guide the face generation process. In the following, we first give a brief overview of problem statement and the proposed ID-GAN. Then, the three players are described in detail.

A. Overview

problem statement. Given an occluded face image $\tilde{\mathbf{X}}$ with an occluded region $\Omega \subset \mathcal{I}$ where $\mathcal{I}$ is the face image region, a face inpainting model $\boldsymbol{I}(\tilde{\mathbf{X}}, \mathbf{M}; \mathbb{W}_I)$ aims to generate an inpainted face image $\mathbf{X}$ by realistically approximating the ground-true face image $\mathbf{Y}$ (generally not available), where $\mathbf{M}$ is a binary mask for labeling the occluded region with 1 inside Ω and $\mathbb{W}_I$ is the set of model parameters. Meanwhile, $\mathbf{X}$ should be well identified by a face recognition model $\boldsymbol{R}(\mathbf{X}; \mathbb{W}_R)$ with a set of parameters $\mathbb{W}_R$ pre-trained for recognizing "normal" faces (*e.g.*, $\mathbf{Y}$). Therefore, the objective is learning an inpainting model $\boldsymbol{I}(\tilde{\mathbf{X}}, \mathbf{M}; \mathbb{W}_I)$ which could generate a photorealistic result $\mathbf{X}$ fitting $\boldsymbol{R}(\mathbf{X}; \mathbb{W}_R)$. In summary, the problem can be formulated as

$$\mathbf{X} = \boldsymbol{I}(\tilde{\mathbf{X}}, \mathbf{M}; \mathbb{W}_I), \tag{1a}$$

$$\mathbf{X} \doteq \mathbf{Y}, \tag{1b}$$

$$\boldsymbol{R}(\mathbf{X}; \mathbb{W}_R) \doteq \boldsymbol{R}(\mathbf{Y}; \mathbb{W}_R), \tag{1c}$$

where $\doteq$ means "equivalence" in some metric. Eq. (1) tries to generate face image (1a) to meet the two rules by reducing the difference between $\mathbf{X}$ and $\mathbf{Y}$ in the views of human perception (1b) and machine perception (1c). Now, the question is how to measure the two rules and what metric are used in a unified framework. Inspired by that, we propose ID-GAN.

ID-GAN. It consists of three players (the generator $\mathbb{G}$, the discriminator $\mathbb{D}$ and the recognizer $\mathbb{R}$) with five networks. $\mathbb{G} = \{\boldsymbol{I}, \boldsymbol{S}\}$ adopts an inpainting network $\boldsymbol{I}$ and a structure network $\boldsymbol{S}$ to reconstruct face images with appearance and structure fidelity. $\mathbb{D} = \{\boldsymbol{D}_g, \boldsymbol{D}_l\}$ contains a global discriminate network $\boldsymbol{D}_g$ and a local discriminate network $\boldsymbol{D}_l$. $\mathbb{R}$ contains a recognition network $\boldsymbol{R}$. The generator takes an occluded face image $\tilde{\mathbf{X}}$ as input and output the inpainted face image $\mathbf{X}$ which preserves appearance and structure fidelity to meet the photorealistic rule. The discriminators classify whether $\mathbf{X}$ is real or fake globally and locally. The recognizer identifies the low-level and high-level semantic features from different network layers between $\mathbf{X}$ and $\mathbf{Y}$ with a collect of identity-centered features, thus meeting the identity-preserving rule. Unlike the existing GAN-based approaches that treat the recognizer as a spectator without competition, the recognizer in our ID-GAN takes part in cooperation with the discriminator as well as competition with the generator via the supervision of identity-centered features. By training on a face set $\mathbb{F} = \{\mathbf{Y}_i, y_i\}_{i=1}^N$, the competition of ID-GAN converges when the generator produces faces that have both high-quality visual appearance and accurate identity clustering characteristics. Here, N is the size of face set and $y_i \in [1, C]$ is the identity label of ith ground-true face image $\mathbf{Y}_i$. After that, the inpainting network $\boldsymbol{I}$ and the recognition network $\boldsymbol{R}$ are cascaded together to recognize occluded faces.

B. Generator $\mathbb{G}$

The generator $\mathbb{G}$ aims to restore the missing region Ω to reconstruct the ground-true face image $\mathbf{Y}$. It takes $\tilde{\mathbf{X}}$ and $\mathbf{M}$ as input and tries to retain the fidelity in appearance and structure. The inpainting network $\boldsymbol{I}$ applies an encoder-decoder architecture with dilated convolution. In this way, appearance fidelity is measured with the spatial discounted reconstruction loss [22]:

$$\mathcal{L}_I = \frac{1}{|\mathcal{I}|} \left\| (\boldsymbol{I}(\mathbf{X}, \mathbf{M}; \mathbb{W}_I) - \mathbf{Y}) \odot \mathbf{W} \right\|^2, \tag{2}$$

where $\|*\|$ is ℓ_2 norm operator and $|\mathcal{I}|$ is the image size, $\mathbf{W}$ is a spatial weighting mask whose pixel value is computed as 0.99^d, where d is the distance of the pixel to the nearest known pixel, and $\odot$ refers to pixel-wise multiplication operator.

Beyond appearance fidelity, another structure network $\boldsymbol{S}$ is introduced for structure fidelity, by requiring the structure consistency between inpainted and ground-true images. We employ a pre-trained semantic parsing model similar to the one in [20] to encourage facial harmony. The structure network functions as a multi-class classifier and assigns a label to every pixel, semantically segmenting the image into k parts (classes), corresponding to different facial features. Then, the structure fidelity is defined as the simple pixel-wise softmax loss:

$$\mathcal{L}_S = -\frac{1}{|\mathcal{I}|} \sum_{i=1}^{|\mathcal{I}|} \log \left(\frac{e^{\mathbf{q}_{i,l_i}}}{\sum_{j=1}^{k} e^{\mathbf{p}_{i,j}}} \right), \tag{3}$$

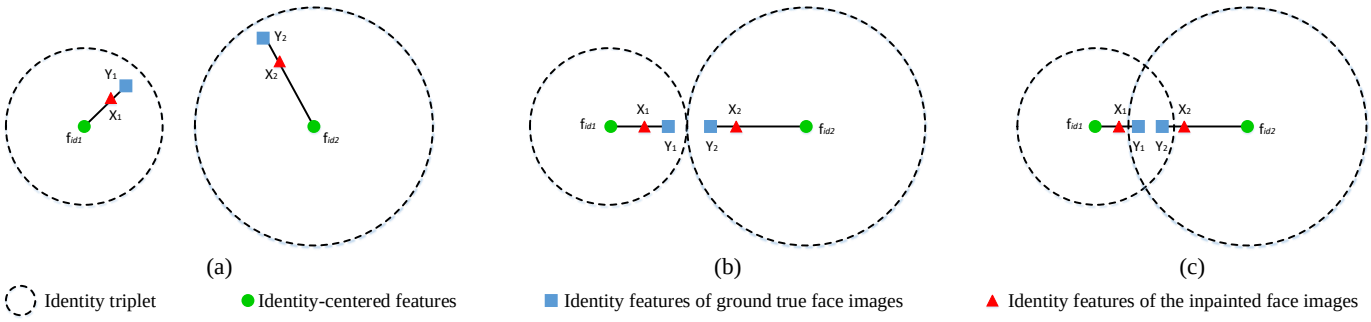

Fig. 3. The merit of the identity-diversity loss. We observe the changes in the separability when the classification surface between two different identity classes gets narrower and find that: (a) the inpainted faces remain being distinguished from each other when identity features of two ground-true face images are separable, (b) the margin is enlarged when identity features of two ground-true face images are very close, and (c) the distance tends to having discriminativeness even when identity features of two ground-true face images are not separable.

where $\mathbf{p} = \boldsymbol{S}(\mathbf{X}; \mathbb{W}_S)$ and $\mathbf{q} = \boldsymbol{S}(\mathbf{Y}; \mathbb{W}_S)$ denote the probability feature maps extracted from $\mathbf{X}$ and $\mathbf{Y}$ by $\boldsymbol{S}$, respectively, $\mathbf{p}_{i,j}$ denotes the predicted probability of the ith pixel belonging to jth class, and l_i is the corresponding ground-true label, having $l_i = \arg\max_j \mathbf{q}_{i,j}$. Here, $\mathbb{W}_S$ is the parameter set of the structure network $\boldsymbol{S}$.

C. Discriminator $\mathbb{D}$

ID-GAN follows the recent approaches [19], [20], [22] and adopts two discriminative networks to preserve global and local textures. The global discriminative network $\boldsymbol{D}_g$ takes the whole image as input, while the local discriminative network $\boldsymbol{D}_l$ uses only the sub-image in Ω. In this way, the contextual information from local and global regions compensates each other, eventually reaching a balance between global consistency and local details. Toward this end, two discriminative networks regularize the inpainting network via local and global adversary loss by solving a min-max optimization problem:

$$\begin{aligned}\mathcal{L}_{D_t} = \min_{\boldsymbol{I}} \max_{\boldsymbol{D}_t} \mathbb{E}[\log \boldsymbol{D}_t(\mathbf{Y}; \mathbb{W}_D^t)) + \\ \log(1 - \boldsymbol{D}_t(\boldsymbol{I}(\mathbf{X}, \mathbf{M}; \mathbb{W}_I); \mathbb{W}_D^t)],\ t \in \{g, l\},\end{aligned} \tag{4}$$

where $\mathbb{W}_D^g$ and $\mathbb{W}_D^l$ are the parameter sets of $\boldsymbol{D}_g$ and $\boldsymbol{D}_l$, respectively. The two networks follow similar architecture in WGAN [52], except that the input is the entire image and local image in Ω respectively. The network architecture consists of ten convolutional layers and a fully-connected layer.

D. Recognizer $\mathbb{R}$

In previous GAN frameworks [23], [22], [40], identity-preservation is usually achieved by enforcing the consistency of semantic features between $\mathbf{X}$ and $\mathbf{Y}$, remembered as perceptual loss [53], [40]. The perceptual loss calculates divergence between semantic features, extracted from a pre-trained deep model which plays the role of a recognizer. Typically, the recognizer in these frameworks is considered as a spectator to regularize $\mathbb{G}$ without competition. By contrast, our ID-GAN treats the recognizer $\mathbb{R}$ as the third player which not only competes with $\mathbb{G}$ but also cooperates with $\mathbb{D}$.

To this end, in addition to the perceptual loss for cooperating with $\mathbb{D}$, ID-GAN proposes an extra identity-diversity loss to guide the competition between $\mathbb{R}$ and $\mathbb{G}$. The identity-diversity loss is measured using a collect of identity-centered features $\mathcal{F} = \{\mathbf{f}_{id}\}_{id=1}^{C}$, where identity-centered feature $\mathbf{f}_{id}$ represents the centroid of high-level semantic features (a.k.a, identity features) for training images with identity label id:

$$\mathbf{f}_{id} = \frac{\sum_{i=1}^{N} \delta(y_i = id) \boldsymbol{R}(\mathbf{Y}_i; \mathbb{W}_R^h)}{\sum_{i=1}^{N} \delta(y_i = id)}, \quad id \in [1, C], \tag{5}$$

where $\delta(y_i = id)$ is an indicator function which equals 1 if $y_i = id$ and 0 otherwise, $\mathbb{W}_R^h$ is the subset of $\mathbb{W}_R$ ranging from the first to the hth layer (*e.g.*, identity layer). The recognizer plays with the other two players via a set of identity triplets $\mathbb{T} = \{(\mathbf{X}_i, \mathbf{Y}_i, \mathbf{f}_{y_i})\}_{i=1}^{N}$. Thus, for an identity triplet $(\mathbf{X}, \mathbf{Y}, \mathbf{f}_{id})$, the recognition loss can be defined as:

$$\begin{aligned}\mathcal{L}_R = \sum_{j=\{l,h\}} \alpha_j \left| \boldsymbol{R}(\mathbf{X}; \mathbb{W}_R^j) - \boldsymbol{R}(\mathbf{Y}; \mathbb{W}_R^j) \right| + \\ \beta \left[\left| \boldsymbol{R}(\mathbf{X}; \mathbb{W}_R^h) - \mathbf{f}_{id} \right| - \left| \boldsymbol{R}(\mathbf{Y}; \mathbb{W}_R^h) - \mathbf{f}_{id} \right|, 0 \right]_+,\end{aligned} \tag{6}$$

where $\mathbb{W}_R^l$ is the subset of $\mathbb{W}_R$ for extracting low-level semantic features. $[*, 0]_+$ stands for $\max$ operator. The balancing constants are set as $\alpha_l = 0.5$, $\alpha_h = 0.05$ and $\beta = 0.5$.

Eq. 6 consists of an instance-wise term and an identity-diversity term. The instance-wise term is a modified version of perceptual loss [53] used for encoding the semantic consistency between the inpainted face image $\mathbf{X}$ and the ground-truth face image $\mathbf{Y}$. It measures the difference of the semantic features from a low-level and a high-level layer of the recognizer. As a result, $\mathbf{X}$ tends to be close to $\mathbf{Y}$ in semantic. The identity-diversity term measures differences of the distances towards the identity center and distinguish the diversity among face images in a same identity class, which forces $\mathbf{X}$ towards its identity center (*e.g.*, $\mathbf{f}_{id}$). While the instance-wise term tries to approach the original faces as possible, the identity-wise term rearranges the distribution of inpainted faces in identity feature space, via enforcing them to have similar clustering characteristics as original ones. In this way, the recognition loss tends to reduce the intra-image difference (between $\mathbf{X}$ and $\mathbf{Y}$) and increase the inter-image difference in the same identity class.

In addition, the setting of Eq. 6 can meet the learning case like [37] that has training data with limited depth (a

relatively small number of samples for each class) and sufficient breadth (many classes). Since the task of face inpainting allows varieties of plausible possibilities, we encourage the diversity of inpainted faces to enable the model from fully exploring the generation space. Toward this end, we adopt a rather loose regularization by enforcing the inpainted faces to be closer to identity center than the original ones. In this way, we allow certain intra-class turbulence in semantic space caused by expression, lightening, occlusions etc, at the base of preserving inter-class discrimination. The inpainted results therefore could be more realistic and sharp as more diverse details can be learned.

Fig. 3 shows the merit of the identity-diversity loss. It is easy to find that: (a) when the faces of two identity classes in real domain are well separated by the recognizer, the inpainted faces remain separable, (b) when the faces of two identity classes in real domain are getting very close, the loss can increase their margin to improve the discrimination and (c) even when the faces in real domain is classified incorrectly, the loss tends to force them being recognizable. Since ID-GAN tends to push the inpainted faces towards their identity center by minimizing the identity-diversity loss, the inpainted faces with a same identity label tend to cluster together, which obviously preserves identity.

E. Total Loss

The total loss function accumulates multiple losses on identity triplet set $\mathbb{T}$ and is formulated as

$$\mathcal{L} = \sum_{\mathbb{T}} \{\mathcal{L}_I + \lambda_{D_l}\mathcal{L}_{D_l} + \lambda_{D_g}\mathcal{L}_{D_g} + \lambda_S\mathcal{L}_S + \lambda_R\mathcal{L}_R\}, \quad (7)$$

where λ_{D_l}, λ_{D_g}, λ_S and λ_R are weights to balance different different losses. Eq. 1 and Eq. 7 explicitly demonstrated how our model meets the two rules. The photorealistic rule in Eq.(1b) is met by minimizing $\mathcal{L}_I$ and $\mathcal{L}_S$ which are measured by appearance fidelity and structure consistency respectively. On the other hand, the identity-preserving rule in Eq.(1c) is met by penalizing the adversary loss measured by realism as well as the recognition loss measured by semantic difference and identity diversity, respectively. Therefore, the stationary point is reached when the generator produces faces that have high visual quality and preserve identity at the same time.

IV. Experiments

In this section, we first introduce the experiment setting, then conduct two experiments, in accord to the two rules, to prove the advantage of the proposed approach by comparing it with 4 state-of-the-arts. In the first experiment, show comprehensive inpainting results of our proposed approach in both qualitative and quantitative. And in the second experiment, we analyze the recognition performance on occluded faces.

A. Experiment Setting

Benchmarks. Considering recognition by inpainting, we benchmark with 4 state-of-the-art inpainting models, including an exemplar-based model **PM** [15], two recent GAN based models **GFC** [20] and **CA** [22], and a recent generative model **CL** [21]. They are summarized in Tab. I. We also denote **GT** and **OCC** as the results generated from ground-true face images and occluded face images, respectively. Our two models **Our-w/oID** and **Our-wID** are trained without and with the identity-diversity loss, respectively.

TABLE I
THE BENCHMARKING APPROACHES

Approach	Idea	Published	Year
PM [15]	Exemplar	ICCV	2009
GFC [20]	GAN	CVPR	2017
CL [21]	Generative	AAAI	2018
CA [22]	GAN	CVPR	2018

Datasets. We use the CelebA dataset [54] for training and LFW dataset [55] for testing in our experiments. The CelebA dataset consists of $202,599$ face images covering $10,177$ subjects. During the training, we use $162,770$ images for training and the remaining $19,867$ images for validation. The images are cropped and resized to 128×128. The LFW dataset consists of $13,233$ images of $5,749$ identities. To generate occluded faces, we take a $m \times m$ (*e.g.*, $m = 48$) mask to randomly cover on the ground-true face images. The images are preprocessed following the training dataset to prepare the test set. Then, the occluded face images are feeded to each model. The total $13,233$ images on LFW are used to benchmark the inpainting results. Among them, $6K$ pairs, including $3K$ positive pairs and $3K$ negative pairs, are selected to evaluate the performance of occluded face verification.

Models. We use two pre-trained deep models for the structure network $\boldsymbol{S}$ and the recognition network $\boldsymbol{R}$. $\boldsymbol{S}$ takes a face parsing model [20] to generate semantic segmentation, which could use to improve the structure alignment by enforcing the structure consistency between the inpainted result and the ground-true image. For the recognizer, we take VGGFace [36] as an example to integrate into ID-GAN. VGGFace achieves a very high accuracy of 98.95% on LFW test set after the faces are well aligned. To verify the general performance of our occluded face recognition approach, we employ four recent face recognizers (CenterLoss [5], SphereFace [30], VGGFace2 [6] and ArcFace [56]) along with VGGFace to measure the recognition accuracy.

Implementation Details. To prevent over-fitting and improve the generalization ability of the model, we do data augmentation that includes flipping, shift and rotation (±15 degrees). During the training, the size of the mask is fixed to be $m = 48$ but the position is random, preventing the model from latching on certain contents only. The masked region of the image is pre-filled with white pixels before inputting into the generator. In this experiment, we set $\lambda_{D_l} = \lambda_{D_g} = 300$, $\lambda_S = 0.05$, and $\lambda_R = 0.0002$. The optimization is implemented in Caffe [57] with ADAM [58] algorithm. The learning rate is set as 0.0001 and $\beta = 0.9$ for generator, and the learning rate is 0.0001 and $\beta = 0.5$ for discriminators.

B. Inpainting Results

First, we investigate the inpainting results in terms of visual quality. We mask the images with 48×48 white blocks and

Fig. 4. Inpainting results generated by different models. In each row, from left to right it is the original face, masked face, inpainting result using **PM** [15], **GFC** [20], **CL** [21], **CA** [22], **Our-w/oID** and **Our-wID** model. From the results, we could see that exemplar-based From the results, we could see that exemplar-based model **PM** may generate artifacts in key facial parts (*e.g.*, eyes) especially when the similar content is no long available in occluded face images, which will heavily hurt the recognition. On the contrast, owing to the introduced identity information in our models, these key parts can be better recovered than those models that take less identity consideration.

Fig. 5. Local zoom-in results from the last row of Fig. 4. Our **Our-wID** model generate more natural eyes, suggesting that the proposed identity-diversity loss facilitates the recovery of the identity-related details.

send it into different inpainting models. Fig.4 shows several examples where faces with a centering mask are inpainted with different models. To be specific, from left to right it is the original face(**GT**), occluded faces(**OCC**), faces inpainted by **PM**, **GFC**, **CL**, **CA**, **Our-w/oID** and **Our-wID** model, respectively. From the results, we could see that exemplar-based model **PM** tend to generate artifacts in key facial parts (*e.g.*, eyes) especially when the similar content is no long available in remaining images. Generative deep models including **GFC** and **CL** learn high-level semantic distribution about faces from large scale of training data, while exhibiting instability without alignment. These resulting artifacts actually heavily hurt the recognition. **CA** model and our two models exhibit the most satisfactory and photo-realistic results. However, as the main optimization target for **CA** is inpainting, it focus on approaching the original unoccluded faces as much as possible, leaving slack regularization in feature space. As shown in Fig. 5, the key parts, in this case, eyes, were recovered as they didn't take enough identity consideration. On the contrast, our **Our-wID** model generate more natural eyes, suggesting that the proposed identity-diversity loss facilitates the recovery of the identity details.

Second, we perform quantitative evaluations with two popular metrics, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). They measure visual quality from the view of appearance and structure. The results in Tab. II show that the inpainting models consistently improve visual quality in both metrics, which agrees with the qualitative results. A related low performance is found by the exemplar-based model **PM**, we can suspect that this model may get trouble in inpainting faces due to the lack of available exemplars, especially when dealing with occlusions of large size. GAN-based models give a high quality result and specially the best one is achieved by our model **Our-wID** which incorporates a well-trained recognizer into GAN to improve the inpainted results from low-level features (*e.g.*, appearance and structure) to high-level attributes (*e.g.*, identity).

C. Recognition Results

After the promising photorealistic results, we further check identity-preservation by evaluating on occluded face recognition. The performance is reported as recognition accuracy measured by VGGFace2 [6], which achives a high accuracy of 99.53%. Here we didn't adopt VGGFace to prevent unequal view. The results are summarized in Fig. 6, where **GT** and

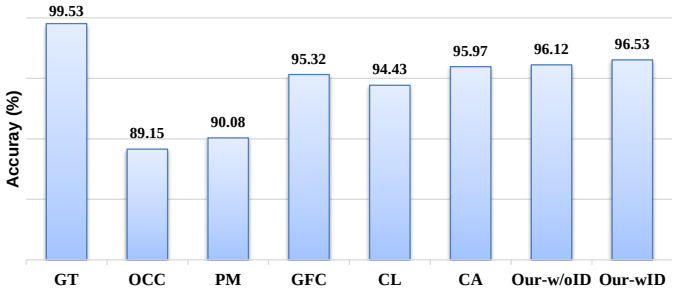

Fig. 6. The accuracy on recognizing occluded faces. The bars denotes the recognition accuracy of ground-truths(**GT**), occluded faces(**OCC**), faces inpainted by **PM**, **GFC**, **CL**, **CA**, **Our-w/oID** and **Our-wID**, respectively.

OCC denotes the models directly used in recognizing ground-true faces and occluded faces, respectively. **OCC** is considered as the baseline model. Fig. 6 demonstrates several important observations. First, occlusion indeed degrade the recognition accuracy sharply (*e.g.*, 10.38% from 99.53% to 89.15% in **OCC**) while face inpainting can typically improve the recognition accuracy more or less, which is as expected.

Second, the traditional exemplar-based approach **PM** achieves a very limited accuracy improvement of 0.93% against the baseline. The reason may come from that its inpainting manner considers less on identity-preservation and thus neglects the introduction of high-level identity knowledge. On the contrast, by explicitly learning to reconstruct some semantic cues from massive face images, the GAN or generative models often reach a reasonably good accuracy.

Finally, to show the impact of the identity-diversity loss, we compare **Our-wID** model with other models as well as our identity-diversity free model **Our-w/oID**. From the results we can find that such improvement is obvious and the model reaches a very high accuracy of 96.53%. Considering large occlusion sizes as well as diverse occluded parts (see Fig. 4), this accuracy implies that the proposed approach provides an efficient way to address occluded faces in the wild. To reveal the reason, we delve into the generated identity features by three better models. To this end, we visualize the features with t-SNE [59] and check the aggregation behaviors in the same identity. The results are shown in Fig. 7 where **Our-wID** achieves most aggregation degree, showing the efficacy of the identity-diversity loss on improving recognition.

TABLE II
QUANTITATIVE PERFORMANCE ON INPAINTING RESULTS.

Metric	OCC	PM	GFC	CL	CA	Our-w/oID	Our-wID
PSNR	14.2216	17.8044	29.6279	25.9527	31.0113	31.1660	31.5588
SSIM	0.8874	0.8912	0.9350	0.8897	0.9565	0.9584	0.9598

TABLE III
RECOGNITION ACCURACY(%) OF INPAINTING RESULTS.

Model	OOC	PM	GFC	CL	CA	Our-tri	Our-cent	Our-w/oID	Our-wID	GT
VGGFace [36]	83.20	83.22	91.05	88.42	90.20	92.63	92.83	92.50	**93.58**	98.95
CenterLoss [5]	74.32	75.45	87.45	83.42	84.65	86.98	87.62	87.77	**88.68**	99.28
SphereFace [30]	75.42	73.67	85.28	84.10	81.55	84.57	85.90	85.00	**86.02**	99.42
VGGFace2 [6]	89.15	90.08	95.32	94.43	95.97	95.52	95.65	96.12	**96.53**	99.53
ArcFace [56]	85.80	84.83	94.83	93.40	91.96	94.47	95.62	95.06	**96.47**	99.82
Average	82.04	81.55	90.69	88.86	88.28	90.83	91.52	91.22	**92.26**	99.40

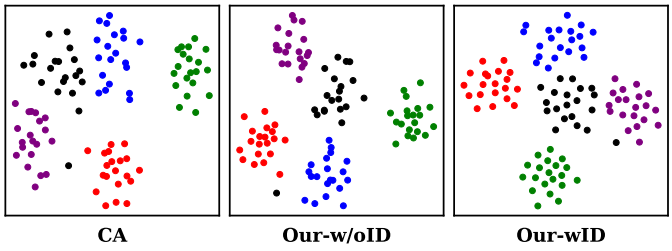

Fig. 7. Visualization of the identity features with t-SNE. Compared to the first picture(**OCC**) where identity features scatter due to occlusion, our **Our-wID** could well aggregate identity features through identity-diversity regularization.

D. Performance Analysis

With improved results achieved by using ID-GAN, we further look at its training manner and analyze the influence of key setting and failure cases.

First, we calculate the recognition accuracy with different recognizers to examine the general performance on occluded face recognition task. Apart from VGGFace, which is adopted as part of ID-GAN, we employ CenterLoss [5], SphereFace [30], VGGFace2 [6] and ArcFace [56] to further demonstrate the identity discriminativeness. The results are shown in Tab.III, where **GT** and **OCC** denotes the recognition accuracy for ground-true faces and occluded faces, **PM**, **GFC**, **CL**, **CA**, **Our-w/oID** and **Our-wID** denote accuracy for inpainted faces generated by these corresponding inpainting models, respectively. And the bottom row is the average accuracy with these five different recognizers. From the table, we can notice that in all cases, our proposed **ID-GAN** achieve the highest recognition accuracy. This prove that our identity-diversity loss, instead of simply fitting the inpainting results relative to the identity regularizer, indeed help the inpainted faces to preserve similar structure in identity space and therefore, facilitate the task of recognition.

Then, we investigate the effect of triplet training with inpainted faces as instances. We extract all triplets from the training dataset, each containing a sample to be inpainted, one positive sample with the same identity and one negative sample with different identity. During the training process, all three images in the triplet are occluded then inpainted by the generator $\mathbb{G}$. We follow experiment settings of two discriminators and parsing network, only replacing the identity-diversity loss with triplet loss. The results are shown in Tab.III (denoted as **Our-tri**). We achieve an accuracy of 92.63% on VGGFace, with minor improvements than **Our-w/oID**, which is trained without identity concerning loss. For other recognizers that are different from the one adopted in training pipeline, **Our-tri** presents lower accuracy than **Our-w/oID**. This implies that the improvements in feature space rearrangement via direct triplet training is somehow limited and hardly general. We suspect the main reason comes from that triplet training is difficult to converge and can easily get over-fitting. Though triplet training echoes with our intend to increase the inter-class divergencies and decrease the intra-class differences, it is not as efficient.

We next study the effect of directly pushing the inpainted faces to the class center. In this experiment, we replace the identity-diversity loss with distance regularization (here we adopt l2 loss) between the inpainted faces and identity center. From the results in Tab.III (denoted as **Our-cent**), it achieves an accuracy of 92.83% on LFW benchmark, measured by VGGFace. As for other recognizers, the increasement is minor even negative (when recognized by CenterLoss and SphereFace). We suspect it is because that, during the training process, the goal of optimization is to find the tradeoff between visual satisfactory and feature consistency. Strict regularization like center loss contracts the optimization for identity feature space, while making satisfactory visual convergence difficult. During the training. It is not as stable as our model which adopts a rather loose constraint instead. And this eventually leads to difficulty in converging to global minima.

In the last experiment, we demonstrate some failure cases of ID-GAN in Fig. 8. We present two positive pairs (in green rectangles at the upper row) and two negative pairs (in blue rectangle at the bottom row), where recognizers fail to give correct estimations even the inpainted results show visually natural appearances. From the figure, two conclusions may be drawn. First, the inpainting model may tend to generate confusing contents under unsatisfactory imaging condition such as illumination and photographing angle. This can be

Fig. 8. Failure examples of face verification after inpainting. Two positive pairs (in green rectangles) are identified as different identities, while two negative pairs (in blue rectangles) are identified as the same person.

explained by the fact that the training CelebA dataset we employed is mostly composed of high-quality front faces of celebrities. When test on low-quality images, the domain gap lied between training and testing data brings in difficulty for the inpainting model to generate perfect results. One probable solution to address this limitation may be further fine-tuning, or directly training the model on datasets with more quality diversity. Second, the negative pair at right-bottom corner is hardly similar in human eyes, while achieving high similarity for recognizers. We suspect it is due to the gap between visual and semantic spaces.

V. Conclusion

In this work, we propose an approach for facilitating the capacity of well-trained face recognizers on identifying occluded faces by inpainting with an identity-diversity GAN (ID-GAN). In ID-GAN, the recognizer is treated as the third player to compete with the generator, leading to high quality and identity-preserving results. In particular, an identity-diversity loss is proposed to enhance the discriminative capacity on face diversity within the same identity, enforcing the inpainted faces clustering towards their identity centers. The proposed approach is extensively evaluated in both visual quality and identity-preservation on popular benchmark dataset. We conduct ablation studies and comparisons with 4 state-of-the-arts, proving the model's efficacy on recognizing occluded faces in the wild. In the future, we plan to extend the framework and the identity-diversity idea to other applications including image-to-image translation and low-resolution recognition.

References

[1] S. Ge, J. Li, Q. Ye, and Z. Luo, "Detecting masked faces in the wild with lle-cnns," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2682–2690.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1701–1708.

[3] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*, 2016, pp. 499–515.

[6] Q. Cao, L. Shen, W. Xie, and *et al.*, "Vggface2: A dataset for recognising faces across pose and age," in *IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 67–74.

[7] J. Zhao, L. Xiong, J. Li, and *et al.*, "3d-aided dual-agent gans for unconstrained face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2018.

[8] J. Wright, A. Y. Yang, A. Ganesh, and *et al.*, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[9] M. Yang, L. Zhang, J. Yang, and D. Zhang, "Robust sparse coding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 625–632.

[10] J. Qian, J. Yang, F. Zhang, and Z. Lin, "Robust low-rank regularized regression for face recognition with occlusion," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 21–26.

[11] H. Li and C. Y. Suen, "Robust face recognition based on dynamic rank representation," *Pattern Recognition*, vol. 60, pp. 13–24, 2016.

[12] X. Wei, C.-T. Li, Z. Lei, and *et al.*, "Dynamic image-to-class warping for occluded face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2035–2050, 2014.

[13] C. Xiong, X. Zhao, D. Tang, and *et al.*, "Conditional convolutional neural network for modality-aware face recognition," in *IEEE International Conference on Computer Vision*, 2015, pp. 3667–3675.

[14] Y. Deng, Q. Dai, and Z. Zhang, "Graph laplace for occluded face completion and recognition," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2329–2338, 2011.

[15] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics*, vol. 28, no. 3, p. 24, 2009.

[16] K. He and J. Sun, "Image completion approaches using the statistics of similar patches," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 12, pp. 2423–2435, 2014.

[17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, and *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[18] D. Pathak, P. Krahenbuhl, J. Donahue, and *et al.*, "Context encoders: Feature learning by inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[19] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 107:1–14, 2017.

[20] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3911–3919.

[21] J. Xie, Y. Lu, R. Gao, and Y. N. Wu, "Cooperative learning of energy-based model and latent variable model via mcmc teaching," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 4292–4301.

[22] J. Yu, Z. Lin, J. Yang, and *et al.*, "Generative image inpainting with contextual attention," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5505–5514.

[23] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 3, pp. 637–647, 2017.

[24] S. Chakraborty, S. K. Singh, and P. Chakraborty, "Local gradient hexa pattern: A descriptor for face recognition and retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 1, pp. 171–180, 2018.

[25] C. Low, A. B. J. Teoh, and C. J. Ng, "Multi-fold gabor, pca, and ICA filter convolution descriptor for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 115–129, 2019.

[26] J. Hu, J. Lu, Y. Tan, and *et al.*, "Local large-margin multi-metric learning for face and kinship verification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1875–1891, 2018.

[27] L. Wang, H. Wu, and C. Pan, "Manifold regularized local sparse representation for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 4, pp. 651–659, 2014.

[28] Z. Lei, D. Yi, and S. Z. Li, "Learning stacked image descriptor for face recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1685–1696, 2015.

[29] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *Proceedings of the International Conference on Machine Learning*, vol. 2, no. 3, 2016, p. 7.

[30] W. Liu, Y. Wen, Z. Yu, and *et al.*, "Sphereface: Deep hypersphere embedding for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 212–220.
[31] X. Zhang, Z. Fang, Y. Wen, and *et al.*, "Range loss for deep face recognition with long-tailed training data," in *IEEE International Conference on Computer Vision*, 2017, pp. 5409–5418.
[32] H. Wang, Y. Wang, Z. Zhou, and *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.
[33] Y. Zheng, D. K. Pal, and M. Savvides, "Ring loss: Convex feature normalization for face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5089–5097.
[34] C. Xiong, L. Liu, X. Zhao, and *et al.*, "Convolutional fusion network for face verification in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 517–528, 2015.
[35] J. Zhao, J. Han, and L. Shao, "Unconstrained face recognition using a set-to-set distance measure on deep learned features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2679–2689, 2018.
[36] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *Proceedings of the British Machine Vision Conference*, vol. 1, no. 3, 2015, p. 6.
[37] X. Zhu, H. Liu, Z. Lei, and *et al.*, "Large-scale bisample learning on id versus spot face recognition," *International Journal of Computer Vision*, vol. 127, pp. 684–700, 2019.
[38] D. Jin and X. Bai, "Patch-sparsity-based image inpainting through a facet deduced directional derivative," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 5, pp. 1310–1324, 2018.
[39] J. Sulam and M. Elad, "Large inpainting of face images with trainlets," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1839–1843, 2016.
[40] C. Wang, C. Xu, C. Wang, and D. Tao, "Perceptual adversarial networks for image-to-image transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 4066–4079, 2018.
[41] F. Zhao, J. Feng, J. Zhao, and *et al.*, "Robust lstm-autoencoders for face de-occlusion in the wild," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 778–790, 2017.
[42] T. Berg and P. N. Belhumeur, "Tom-vs-pete classifiers and identity-preserving alignment for face verification." in *Proceedings of the British Machine Vision Conference*, 2012, pp. 1–11.
[43] Z. Zhu, P. Luo, X. Wang, and X. Tang, "Deep learning identity-preserving face space," in *IEEE International Conference on Computer Vision*, 2013, pp. 113–120.
[44] F. Shiri, F. Porikli, R. Hartley, and P. Koniusz, "Identity-preserving face recovery from portraits," in *IEEE Winter Conference on Applications of Computer Vision*, 2018, pp. 102–111.
[45] B. Dolhansky and C. Canton Ferrer, "Eye in-painting with exemplar generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7902–7911.
[46] Y. Shen, P. Luo, J. Yan, and *et al.*, "FaceID-GAN: Learning a symmetry three-player gan for identity-preserving face synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 821–830.
[47] J. Bao, D. Chen, F. Wen, and *et al.*, "Towards open-set identity preserving face synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6713–6722.
[48] N. Wang, X. Gao, L. Sun, and J. Li, "Anchored neighborhood index for face sketch synthesis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2154–2163, 2018.
[49] Y. Li, L. Song, X. Wu, and *et al.*, "Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 7057–7064.
[50] X. Shu, J. Tang, H. Lai, and *et al.*, "Personalized age progression with aging dictionary," in *International Conference on Computer Vision*, 2015, pp. 3970–3978.
[51] X. Shu, J. Tang, Z. Li, and *et al.*, "Personalized age progression with bi-level aging dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 905–917, 2018.
[52] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.
[53] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 694–711.
[54] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.
[55] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.
[56] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.
[57] Y. Jia, E. Shelhamer, J. Donahue, and *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Internation Coference on Multimedia*, 2014, pp. 675–678.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
[59] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

Shiming Ge (M'13-SM'15) is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He is also the member of Youth Innovation Promotion Association, Chinese Academy of Sciences. Prior to that, he was a senior researcher and project manager in Shanda Innovations, a researcher in Samsung Electronics and Nokia Research Center. He received the B.S. and Ph.D. degrees both in Electronic Engineering from the University of Science and Technology of China (USTC) in 2003 and 2008, respectively. His research mainly focuses on computer vision, data analysis, machine learning and AI security, especially efficient learning models and solutions toward scalable applications.

Chenyu Li is currently a Ph.D. candidate at the Institute of Information Engineering at Chinese Academy of Sciences and the School of Cyber Security at the University of Chinese Academy of Sciences. She received the B.S. degree from the School of Electronics and Information Engineering at the Tongji University. Her research interests are computer vision and deep learning.

Shengwei Zhao received his B.S. degree from the School of Mathematics and Statistics in Wuhan University in 2017. He is now a Master student at the Institute of Information Engineering at Chinese Academy of Sciences and the School of Cyber Security at the University of Chinese Academy of Sciences. His major research interests are deep learning and computer vision.

Dan Zeng received her Ph.D. degree in circuits and systems in 2008, and her B.S. degree in electronic science and technology in 2003, both from University of Science and Technology of China, Hefei. She is currently a full professor at the Key Laboratory of Specialty Fiber Optics and Optical Access Networks and Shanghai Institute of Advanced Communication and Data Science, Shanghai University, Shanghai. Her research interests include computer vision, multimedia content analysis and machine learning.