Pattern Recognition 124 (2022) 108465

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Contrastive attention network with dense field estimation for face completion

Xin Ma[a,b,1,*], Xiaoqiang Zhou[b,d,1], Huaibo Huang[a,b], Gengyun Jia[a,b], Zhenhua Chai[c], Xiaolin Wei[c]

[a] *School of Artificial Intelligence, University of Chinese Academy of Sciences, China*
[b] *NLPR & CEBSIT & CRIPAC, CASIA, China*
[c] *Visual Intelligence Department, Meituan, China*
[d] *University of Science and Technology of China, China*

ARTICLE INFO

Article history:
Received 23 January 2021
Revised 20 November 2021
Accepted 26 November 2021
Available online 27 November 2021

2010 MSC:
00-01
99-00

Keywords:
Face completion
Unsupervised learning
Attention mechanism
3D Face analysis

ABSTRACT

Most modern face completion approaches adopt an autoencoder or its variants to restore missing regions in face images. Encoders are often utilized to learn powerful representations that play an important role in meeting the challenges of sophisticated learning tasks. Specifically, various kinds of masks are often presented in face images in the wild, forming complex patterns, especially in this hard period of COVID-19. It's difficult for encoders to capture such powerful representations under this complex situation. To address this challenge, we propose a self-supervised Siamese inference network to improve the generalization and robustness of encoders. It can encode contextual semantics from full-resolution images and obtain more discriminative representations. To deal with geometric variations of face images, a dense correspondence field is integrated into the network. We further propose a multi-scale decoder with a novel dual attention fusion module (DAF), which can combine the restored and known regions in an adaptive manner. This multi-scale architecture is beneficial for the decoder to utilize discriminative representations learned from encoders into images. Extensive experiments clearly demonstrate that the proposed approach not only achieves more appealing results compared with state-of-the-art methods but also improves the performance of masked face recognition dramatically.

1. Introduction

Face completion (a.k.a face inpainting or face hole-filling) aims at filling missing regions of a face image with plausible contents [1]. It is more difficult than general image inpainting because there are high-level identity information, pose variations, etc in face images. Face completion is a fundamental low-level vision task and can be applied to many downstream applications, such as photo editing and face verification [2,3,86]. The target of face completion is to produce semantically meaningful content and reasonable structure information in missing areas.

There are many attempts for face completion, but they usually treat it as a general image inpainting problem. Traditional image inpainting methods [3,8,9] (e.g., PatchMatch) assume that the content to be filled comes from the background area. Therefore, they gradually synthesize plausible stationary contents by copying and pasting similar patches from known areas. The performances of these methods are satisfying when dealing with background inpainting tasks. But non-repetitive and complicated scenes, such as faces and objects, are the Waterloo of these traditional methods because of the limited ability to capture high-level semantics. Recently, deep convolutional neural networks (CNNs) have made great progress in many computer vision tasks [13,87,88,95]. Thus, many deep learning-based methods have been proposed. Benefiting from the powerful ability of representation learning of CNNs, their performance has been significantly improved. These approaches adopt autoencoder or its variant architectures jointly trained with generative adversarial networks (GANs) to hallucinate semantically plausible contents in missing regions [2,14,15]. But these methods still suffer from three problems:

Firstly, various kinds of masks are often presented in face images in the wild, especially in this tough period of COVID-19, which greatly increases the difficulty of image inpainting. Previous image inpainting approaches usually train an encoder and a decoder

[*] Corresponding author.
E-mail addresses: xin.ma@cripac.ia.ac.cn (X. Ma), xq525@mail.ustc.edu.cn (X. Zhou), huaibo.huang@cripac.ia.ac.cn (H. Huang), gengyun.jia@cripac.ia.ac.cn (G. Jia), chaizhenhua@meituan.com (Z. Chai), weixiaolin02@meituan.com (X. Wei).
[1] Xin Ma and Xiaoqiang Zhou have contributed equally to the work.

https://doi.org/10.1016/j.patcog.2021.108465
0031-3203/© 2021 Elsevier Ltd. All rights reserved.

jointly with some commonly-used loss functions (e.g., reconstruction loss, style loss, etc). But encoders still struggle to learn powerful representations from images with various kinds of masks. As a result, these CNN-based approaches will produce unsatisfactory results with obvious artifacts. A naive solution is to design a very deep network to obtain a large model capacity for learning powerful representations. However, it will increase the computational cost heavily and may not help to learn accurate latent representations.

To cope with this limitation, we propose a self-supervised Siamese inference network with contrastive learning. We assume that two identical images with different masks form a positive pair while a negative pair consists of two different images. Contrastive learning aims to maximize (minimize) the similarities of positive pairs (negative pairs) in a representation space. As explored in [16,17], contrastive learning can be regarded as training an encoder to perform a dictionary $look-up$ task. An encoded "query" should be matched with its corresponding "key" (token) and different from others. The "keys" (tokens) in the dictionary are usually sampled from images, patches, or other data types. In order to acquire a large and consistent dictionary, we design a queue dictionary and a momentum-updated key encoder. As demonstrated in MoCo [16], the proposed self-supervised inference network can learn good features from input images. Thus, the robustness and the accuracy of the encoder can be improved.

Secondly, previous methods consider image inpainting as a conditional image generation task. The roles of the encoder and decoder are recognizing high-level semantic information and synthesizing low-level textures [18], respectively. These approaches, e.g., PConv [15] and LBAM [14], focus more on missing areas and synthesize realistic alternative contents by a well-designed architecture or some commonly-used loss functions. However, there are either obvious color contrasts or artificial edge responses, especially in the boundaries of results produced by these methods since they ignore the structural consistency. In fact, the development of biology has revealed that the human visual system is more sensitive to the topological distinction [19]. Therefore, we focus not only on the structural continuity of restored images surrounding holes but also on generating texture-rich images.

To properly suppress color discrepancy and artifacts in boundaries, we propose a novel dual attention fusion module (DAF) to synthesize pixel-wise smooth contents, which can be inserted into autoencoder architectures in a plug-and-play way. The core idea of the fusion module is to calculate the similarity between the synthesized content and the known region. Some methods are proposed to address this problem, such as DFNet [20] and Perez's method [21]. However, these methods lack flexibility in handling different information types (e.g., different semantics), hindering learning more discriminative representations. Our proposed DAF is developed to adaptively recalibrate channel-wise features by taking interdependencies between channels into account and force CNNs to focus more on unknown regions. DAF will predict an adaptive spatial attention map to blend restored contents and original images naturally.

Finally, the verification performance heavily relies on the pixel level similarity and feature level similarity according to Zhang et al. [22], which means that the geometric information of the output results should be similar to the input. In practice, face appearance will be influenced by a number of factors such as meshes, wearing masks [22–24] and so on. Masks can significantly destroy the facial shape and geometric information, greatly increasing the difficulty of generating visually appealing results. Therefore, it inevitably leads to a sharp decline in face verification performance. For example, healthcare workers must wear sanitary masks to avoid infection of diseases, and they will fail to pass through the face verification system.

In this paper, we assume that the geometric information of the input face image should be kept intact. Inspired by recent advances in 3D face analysis [25,26], a dense correspondence field estimation is integrated into our network since it contains the complete geometric information of the input face. For simplicity, instead of using another network to predict the dense correspondence field separately, we make our decoder simultaneously predict the dense correspondence field and feature maps at multi-scales. Thus, we subtly employ a 3D supervision for our network provided by the dense correspondence field. Under this 3D geometric supervision, our network can generate inpainting results with reasonable structure information.

Qualitative and quantitative experiments are conducted on multiple datasets to evaluate our proposed method. The experimental results demonstrate that our proposed method not only outperforms state-of-the-art methods in generating high-quality inpainting results but also improves the performance of masked face recognition dramatically.

This paper is an extension of our previous conference publication [27]. We extend it in three folds: 1) A dense correspondence field is proposed to be integrated into our network for utilizing 3D prior information of human faces. It can help our network to retain the facial shape and appearance information from the input. 2) We mainly concentrate on face image completion rather than other types of images. We add an extra face dataset, Flickr-Faces-HQ (FFHQ) [28], to demonstrate the effectiveness of our method. 3) We conduct an identity verification evaluation for face completion. It clearly shows the advantage of the proposed method compared with state-of-the-art methods.

To sum up, the main contributions of this paper are as follows:

- We propose a Siamese inference network based on contrastive learning for face completion. It helps to improve the robustness and accuracy of representation learning for complex mask patterns.
- We propose a novel dual attention fusion module that can explore feature interdependencies in spatial and channel dimensions and blend features in missing regions and known regions naturally. Smooth contents with rich texture information can be naturally synthesized.
- To keep structural information of the input intact, the dense correspondence field that binds 2D and 3D surface spaces is estimated in our network, which can preserve the expression and pose of the input.
- Our proposed method achieves smooth inpainting results with rich texture and reasonable topological structural information on three standard datasets against state-of-the-art methods, and also greatly improves the performance of face verification.

2. Related work

2.1. Image inpainting

Image inpainting aims to generate alternative contents when a given image is partially occluded or corrupt. Early traditional image inpainting methods are mainly diffusion-based [1] or patch-based [3]. They often use the information of the pixels (or image patches) around the occluded area to fill the missing regions. Bertalmio et al. [1] proposed an algorithm to fill missing regions with information surrounding them automatically based on the principle that isophote lines arriving at the boundaries of the regions are completed inside. Barnes et a. [3] presented a fast nearest neighbor searching algorithm named PatchMatch, to search and paste the most similar image patches from the known regions. These methods utilize low-level image features to guide the feature propagation from known image backgrounds or image datasets to corrupted regions. Criminisi et al. [29] proposed an efficient algorithm, which combined the advantages of 'texture synthesis' techniques and 'inpainting' techniques. Specifically, they designed a best-first method to find the most similar patches and used them to recover the corrupted regions gradually. These methods work well when holes are small and narrow, or there are plausible matching patches in uncorrupted regions. However, when suffering from complicated scenes, it is difficult for these approaches to produce semantically plausible solutions, due to a lack of semantic understanding of images.

Nowadays, deep learning techniques have made great contributions to computer vision communities. In order to accurately recover corrupted images, many methods adopt deep convolutional neural networks (CNNs) [93,96], especially generative adversarial networks (GANs) [35] in image inpainting. Pathak et al. [30] formulates image inpainting as a conditional image generation problem. Then, they proposed a Context Encoder to recover corrupted regions according to surrounding pixels. Iizuka et al. [36] utilized two discriminators to improve the quality of the generated images at different scales, facilitating both globally and locally consistent image completion. At the same time, some approaches designed a coarse-to-fine framework to solve the sub-problem of image inpainting in different stages [31,37,38]. Nazeri et al. [37] proposed to firstly recover the edge map of the corrupted image, then generate image textures in the second stage. Ren et al. [38] proposed a method in which a structure reconstructor was employed to generate the missing structures of the inputs while a texture generator yielded image details. Zhang et al. [39] proposed an iterative inpainting approach that contained a corresponding confidence map in results. They used this map as feedback and recovered holes by trusting high-confidence pixels.

As a branch of image inpainting, face completion is different from general image inpainting since its target mainly focuses on restoring the topological structure and texture of the face input. Zhang et al. [22] argued that the performance of verification relied on both the pixel level similarity and the feature level similarity. Therefore, they proposed a feature-oriented blind face inpainting framework. Cai et al. [40] proposed a method named FCSR-GAN to perform face completion and face super-resolution by multi-task learning where the generator was required to generate a high-resolution face image without occlusion from the occluded low-resolution face image. Zhou et al. [41] argued that previous works overlooked the serious impacts of inaccurate attention scores. Thus, they integrated the oracle supervision signal into the attention module to produce reasonable attention scores.

2.2. Unsupervised representation learning

Unsupervised learning has shown great potential to learn powerful representations of images recently [16,42,43]. Compared with supervised learning, unsupervised learning utilizes unlabeled data to learn representations, which can go back to as far as the literature proposed by Becker and Hinton [44]. Dosovitskiy et al. [45] proposed to discriminate between a set of surrogate classes generated by applying a number of transformations. Wu et al. [46] treated instance-level discrimination as a metric learning problem. Then, the discrete memory bank was utilized to store the features for each instance. Zhuang et al. [47] maximized a dynamic aggregation metric, which can move similar data instances together in the embedding space and separate dissimilar instances. He et al. [16] proposed a dynamic dictionary consisting of a queue encoder and a moving-averaged encoder from a perspective on contrastive learning and they called this method MoCo. At the same time, Chen et al. [43] also presented a simple framework with contrastive learning for visual representations (SimCLR). Technically, they simplified recent contrastive learning-based algorithms and did not require specific structures and memory banks. Unsupervised learning strategies are also used in many computer vision tasks recently. Mustikovela et al. [48] used self-supervised learning for viewpoint estimation by making use of generative consistency and symmetry constraint. Zhan et al. [49] utilized a mask completion network to predict occlusion ordering with a self-supervised learning strategy.

2.3. Attention mechanism

Attention mechanism is a hot topic in computer vision and has been widely investigated in many works [50,89–91]. The wildly-used attention mechanism can be coarsely divided into two categories: spatial attention [50] and channel attention [53] for image inpainting. Yu et al. [31] argued that convolutional neural networks lacked the ability to borrow or copy information from distant places, which led to blurry textures in generated images. Thus, they proposed a contextual attention module to calculate the spatial attention scores between pixels in the corrupted region and known region. Hong et al. [20] proposed a fusion block to generate an adaptive spatial attention map α to combine features in the corrupted region and known region. In this paper, we investigate both spatial attention and channel attention mechanism to further improve the performance of face completion.

2.4. 3D Face analysis

Nowadays, the famous 3DMM [54] is widely used to express facial shape and appearance information for face related tasks, such as facial attribute editing, face hallucination, etc [55,56]. Roth et al. [57] proposed a photometric stereo-based method for unconstrained 3D face reconstruction, which benefited from a combination of landmark constraints and photometric stereo-based normals. Yin et al. [45] proposed a generative adversarial network combined with 3DMM, termed as FF-GAN, to provide shape and appearance priors without requiring large training data. 2DASL [58] utilized 2D face images with noisy landmark information in the wild to assist 3D face model learning. It has become a popular method to establish the dense correspondence field between the 2D and 3D space. Gōler et al. [12,25,26] proposed a UV correspondence field to build pixel-wise correspondence between RGB color space and 3D surface space. These works show that the UV correspondence field can retain geometric information of the human face.

3. Methodology

In this section, we first present our self-supervised Siamese inference network. Subsequently, the details of the dual attention fusion (DAF) module, the dense correspondence estimation, and learning objectives in our method are provided. The overall framework of our face completion method is shown in Fig. 1.

3.1. Self-supervised Siamese inference network

Our proposed self-supervised Siamese inference network consists of two identical encoders but not sharing parameters [16,17,59], noted as E_q and E_k, respectively. The proposed inference network is trained with contrastive learning, which can be viewed as training an encoder to perform a dictionary look-up task: a 'query' encoded by E_q should be similar with its corresponding 'key' (i.e., positive key) represented by another encoder E_k and dissimilar to others (i.e., negative keys). Two images with different masks are required for the proposed inference network, named as x_q and x_k, respectively. Thus, we can obtain a query representation $z_q = E_q(x_q)$ and a key representation $z_k = E_k(x_k)$, respectively.

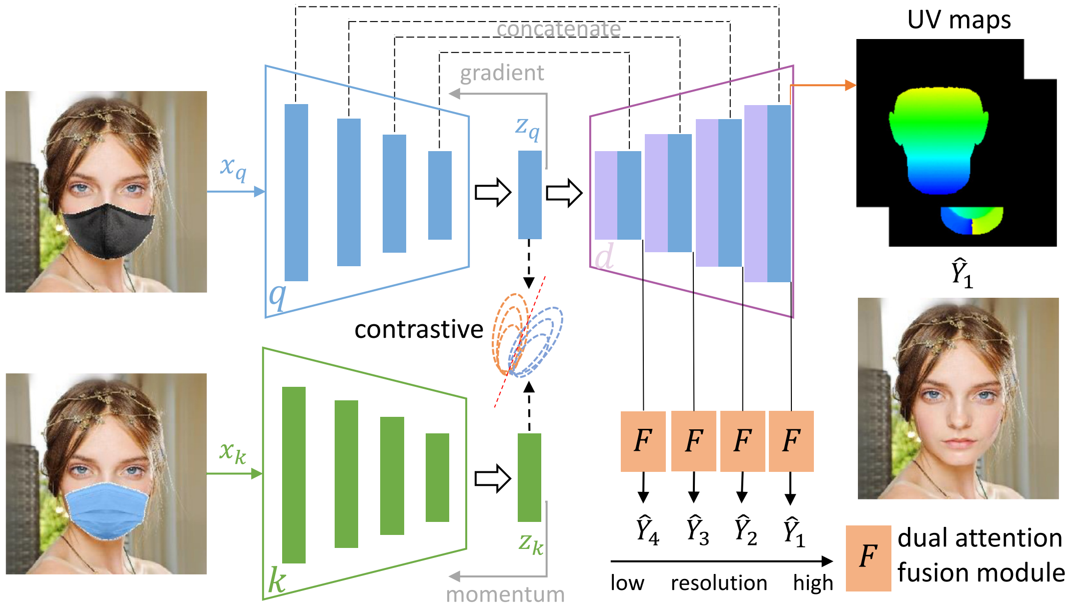

Fig. 1. The network architecture of our method. The self-supervised Siamese inference network consists of encoders E_q and E_k. This inference network encodes the new key representations on-the-fly by using the momentum-updated encoder E_k. We insert the dual attention fusion module into several decoder layers, forming a multi-scale decoder. We allow the decoder to estimate the dense correspondence field and the feature maps that are used for the DAF module at multi-scales simultaneously. The inference network is firstly trained with contrastive learning. Then the pre-trained encoder E_q and the decoder are jointly trained with the fusion module.

Following many previous self-supervised works [47,60], the contrastive loss is utilized as the self-supervised objective function for training the proposed inference network and can be written as:

$$\mathcal{L} = -log\frac{exp(z_q.z_k^+/\tau)}{\sum_{i=0}^{K} exp(z_q.z_{k_i}/\tau)}, \tag{1}$$

where τ is the temperature hyper-parameter, and the loss function will degrade into the original $softmax$ when τ is equal to 1. The output will be less sparse with τ increasing [61]. The τ is set as 0.07 for the efficient training process in this work. Specially, this loss, also known as InfoNCE loss [16,17], tries to classify z_q as z_k^+. Here, z_q and z_k^+ are encoded from a positive pair. K means the number of negative samples.

High-dimensional continuous images can be projected into a discrete dictionary by contrastive learning. There are three general mechanisms for implementing contrastive learning (i.e., end-to-end training [17], memory bank [46] and momentum updating [16]), whose main differences are how to maintain keys and how to update the key encoder. Considering GPU memory size and powerful feature learning, we follow MoCo [16] to design a consistent dictionary implemented by *queue*. Thus, the key representations of the current batch data are enqueued into the dictionary while the oldest representations are dequeued progressively. The length of the queue is under control, which enables the dictionary to contain a large number of negative image pairs. Such a dictionary with large-scale negative pairs will facilitate representation learning. We set the length of the queue as 65536 in this work.

It is worth noting that the encoder E_k is updated by a momentum-updated strategy instead of direct back-propagation. The main reason is that it's difficult to propagate the gradients to all keys in the queue. The updating process of E_k can be formulated as follows:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \tag{2}$$

where θ_q and θ_k denotes as the parameters of E_q and E_k, respectively. θ_q is updated by back-propagation. $m \in [0,1)$ is the momentum coefficient hyper-parameter and set as 0.9 in this paper. The momentum-update mechanism makes the encoder E_k update smoothly relative to E_q, resulting in a more consistent discrete dictionary.

3.2. Dual attention fusion module

We now give more details about our proposed dual attention fusion module (see Fig. 2), which contains a channel attention mechanism and a spatial attention mechanism. This fusion module is embedded into the last several layers of the decoder and outputs face completion results with multi-scale resolutions [62]. Thus, constraints can be imposed on multi-scale outputs for high-quality results.

Previous CNN-based image inpainting approaches treat channel-wise features equally, thus hindering the ability of the representation learning of the network. Meanwhile, high-level and inter-related channel features can be considered as specific class responses. For more discriminative representations, we first build a channel attention module in our proposed fusion module.

As shown in Fig. 2, let a feature map $F = [f_1, \cdots, f_c, \cdots, f_C]$ be one of the inputs of the fusion module, whose channel index is c and size is $h \times w$. The channel descriptor can be acquired from the channel-wise global spatial information by global averaging pooling. Then we can obtain the channel-wise statistics $z_c \in \mathbb{R}^c$ by shrinking F:

$$z_c = H_{GP}(f_c) = \frac{1}{h \times w}\sum_{i=1}^{h}\sum_{j=1}^{w} f_c(i,j), \tag{3}$$

where z_c is the cth element of z. $f_c(i,j)$ is the value at position (i,j) of cth feature f_c. H_{GP} means the global pooling function.

In order to fully explore the channel-wise dependencies of the aggregated information, we introduce a gating mechanism. As illustrated in [53,64], the sigmoid function can be used as a gating function:

$$\omega = \sigma(W_U\delta(W_D z)), \tag{4}$$

where $\sigma(\cdot)$ and $\delta(\cdot)$ are the sigmoid gating and ReLU functions, respectively. W_D and W_U are the weight sets of *Conv* layers who

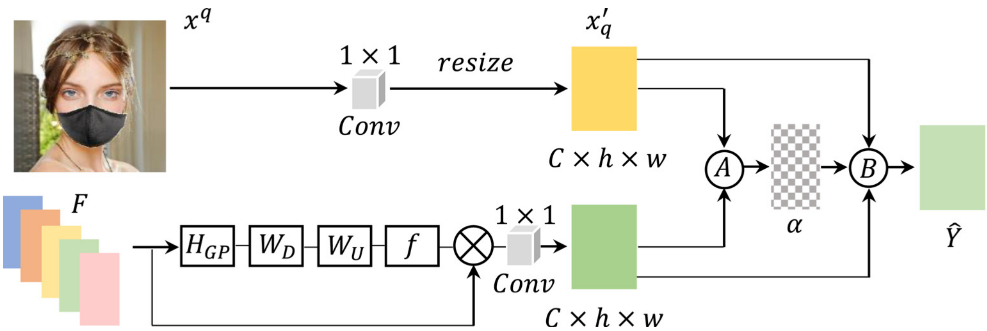

Fig. 2. The architecture of the dual attention fusion module. It firstly predicts an adaptive spatial attention map α with the learnable transformation function $\mathcal{A}$. Then we can obtain final natural face completion results with rich texture by the fusion function $\mathcal{B}$.

set channel number as C/r and C, respectively. Finally, the channel statistics ω are acquired and used to rescale the input f_c:

$$\hat{f}_c = w_c \cdot f_c, \tag{5}$$

where w_c and f_c are the scaling factor and feature map of the cth channel, respectively.

The long-range contextual information is essential for discriminant feature representations. We propose a spatial attention module that forms the final part of the proposed fusion module. Given an input image with a mask x_q, we first get $x_q{}'$ that matches the size of the re-scaled feature map $\hat{F} \in \mathbb{R}^{c\times h\times w}$,

$$x_q{}' = (W_C x_q)\downarrow, \tag{6}$$

where W_C and ↓ are the weight set of a 1 × 1 convolutional layer and downsample module, respectively.

Then the adaptive spatial attention map $\alpha \in \mathbb{R}^{C\times h\times w}$ is given by,

$$\alpha = \sigma(\mathcal{A}(W_K\hat{F}, x_q{}')), \tag{7}$$

where W_K is the weight set of a 1 × 1 convolutional layer. It sets channel number of $\hat{F}$ to be same with $x_q{}'$. $\mathcal{A}$ is a learnable transformation function implemented by three 3 × 3 convolutional layers. $W_K\hat{F}$ and $x_q{}'$ are first concatenated and then fed into the convolutional layers. $f(\cdot)$ is the sigmoid function that can make α an attention map to some extent.

The final inpainting result $\hat{Y}$ is obtained by,

$$\hat{Y} = \mathcal{B}(\alpha, W_K\hat{F}, x_q{}') = \alpha \odot W_K\hat{F} + (1-\alpha)\odot x_q{}', \tag{8}$$

where $\odot$ and $\mathcal{B}$ denote the Hadamard product and fusion function, respectively. The adaptive spatial attention map α can adjust the balance between the ground truth image and the restored image to obtain a smoother transition. We can eliminate obvious color contrasts and artifacts especially in boundary areas, and get natural face completion results with richer textures.

3.3. Dense correspondence field estimation

Masks can dramatically destroy the facial shape and structure information, such as viewing angles, facial expressions, and so on, making it quite tough to achieve visually appealing results. To keep the geometric information of the human face intact during the face completion process, we introduce a dense correspondence field that binds the 2D and 3D surface spaces into our network.

The structure and texture information of the face image can be disentangled by the dense correspondence field according to [25,26]. The geometrical information is stored in the correspondence field while the texture map can represent the surface of a 3D face to some extent. In this paper, we mainly concentrate on inferring the dense correspondence field by our network. Technically, given an input image $x \in \mathbb{R}^{c\times h\times w}$, the dense correspondence field $C = (u; v)$ consists of maps in the UV space ($u, v \in \mathbb{R}^{h\times w}$). The visual illustration is shown in Fig. 3 in which the minimum is rendered as blue and the maximum is rendered as yellow.

We allow our decoder to predict the dense correspondence fields and feature maps at multi-scales simultaneously, where the feature maps are fed into the proposed dual attention fusion module (please see Section 3.2). Thanks to the multi-scale network architecture, our decoder can obtain context information better and maintain geometrical information. In order to supervise C during training, we minimize the pixel-wise error between the estimated result and the ground truth C. It can be written mathematically as,

$$\mathcal{L}_{UV} = ||C' - C||_2, \tag{9}$$

where C' means the predicted dense correspondence field result of an input image. We employ BFM [65], a 3D shape estimation approach, to obtain the ground truth dense correspondence field C similar with [12,55]. We then obtain coordinates of vertices by performing the model fitting method [66]. Finally, those vertices are mapped to the UV space by the cylindrical unwrapping according to Booth and Zafeiriou [67].

3.4. Loss functions

Following [92–94], for synthesizing richer texture details and correct semantics, the element-wise reconstruction loss, the perceptual loss [68], the style loss and the adversarial loss are used in our proposed method. Moreover, we also employ the identity preserving loss function to ensure that the identity information of the generated images remains unchanged.

Reconstruction Loss. It is calculated as $\mathcal{L}_1$-norm between the inpainting result $\hat{Y}$ and the target image Y,

$$\mathcal{L}_{rec} = ||Y - \hat{Y}||_1. \tag{10}$$

Style Loss. For getting richer textures, we also adopt the style loss defined on the feature maps produced by the pre-trained VGG-16. Following [14,15], the style loss can be calculated as the $\mathcal{L}_1$-norm between the Gram matrices of the feature maps,

$$\mathcal{L}_{style} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{C_i\cdot C_i}||\Phi^i(Y)(\Phi^i(Y))^T - \Phi^i(\hat{Y})(\Phi^i(\hat{Y}))^T||_1, \tag{11}$$

where C_i denotes the channel number of the feature map at ith layer in the pre-trained VGG-16.

Identity Preserving Loss. To ensure the generated face images belong to the same identity as the target face images, we adopt LightCNN [5] to extract the features, then use the mean square error to constrain the embedding spaces,

$$\mathcal{L}_{ip} = ||\Psi(Y) - \Psi(\hat{Y})||_2, \tag{12}$$

where Ψ means the pre-trained LightCNN network [5].

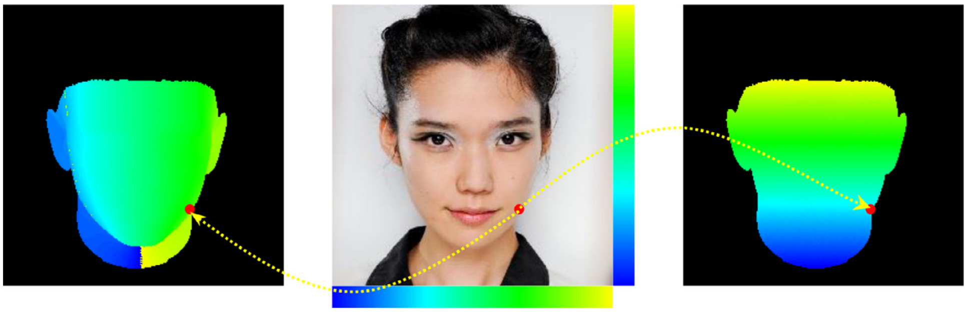

Fig. 3. Visualization examples of the dense correspondence. The face image is shown in the middle. The corresponding U map and V map are shown in the left and right, respectively.

Model Objective. The above loss functions can be grouped into two categories: *Structure Loss* and *Texture Loss*, respectively. The *Structure Loss* is given by,

$$\mathcal{L}_{struct}^{k} = \lambda_{rec}\mathcal{L}_{rec}^{k} + \lambda_{uv}\mathcal{L}_{uv}^{k}, \tag{13}$$

where λ_{rec} and λ_{uv} mean the weight factors and are set as 6 and 0.1 empirically. $\mathcal{L}_{struct}^{k}$ is calculated as the sum of $\mathcal{L}_{rec}$ and $\mathcal{L}_{uv}$ at the kth layer of the decoder. Here, $\mathcal{L}_{uv}$ means the UV loss function (please see Section 3.3).

The *Texture Loss* is given by,

$$\mathcal{L}_{text}^{k} = \lambda_{style}\mathcal{L}_{style}^{k} + \lambda_{ip}\mathcal{L}_{ip}^{k}, \tag{14}$$

where λ_{style} and λ_{ip} are trade-off factors and are set as 240 and 0.1 empirically in this work.

Finally, the total model objective can be formulated as,

$$\mathcal{L}_{total} = \frac{1}{|P|}\sum_{k\in P}\mathcal{L}_{struct}^{k} + \frac{1}{|Q|}\sum_{k\in Q}\mathcal{L}_{text}^{k}, \tag{15}$$

where both P and Q are the selected decoder layer sets that imposed constraints. We select P as $\{1,2,3,4,5,6\}$ and Q as $\{1,2,3\}$ respectively for better inpainting results. Note that 1 represents the outermost layer.

4. Experiments

To demonstrate the superiority of our approach against state-of-the-art methods, both quantitative and qualitative experiments for face completion and face verification experiments are conducted. In this section, we will introduce the details of our experimental settings and the experimental results one by one.

4.1. Datasets and protocols

CelebA. The CelebFaces Attributes dataset [69] is widely used for face hallucination, image-to-image translation, etc. It's a large-scale face attributes dataset containing more than 200k celebrity images, which includes face images with large occlusion and pose variations. We randomly select 10,000 images for testing and the rest for training.

CelebA-HQ. It's a high-resolution face images dataset established by Karras et al. [62], which contains 30,000 high-quality face images. We divide the dataset into two subsets: the training set of 28,000 images and the testing set of 2000 images.

FFHQ. The Flickr-Faces-HQ dataset [28] is a high-quality dataset containing 70,000 face images at 1024 × 1024 resolution. It also covers age, ethnicity, and image background variations. We randomly choose 6000 images for testing and the rest for training.

Multi-PIE. It contains more than 750,000 images that cover 15 viewpoints, 19 illumination conditions and a number of facial expressions of 337 identities [70]. We follow Huang et al. [71] to split the dataset. In our experiments, we only utilize the training set to train our network and the compared methods for face recognition.

LFW. The Labeled Faces in the Wild [72] is a benchmark database commonly used for face recognition, which contains 13,233 images of 5749 people captured in unconstrained environments. LFW provides a standard protocol for face verification that contains 6000 face image pairs (including 3000 positive pairs and 3000 negative pairs, respectively). We use these standard face image pairs to evaluate face verification performance via face completion. Specially, face images in the gallery set remain the same while the counterparts in the probe set are occluded by masks. We firstly recover the occluded face images by our proposed method and the state-of-the-arts. Then, we compare the verification performance. It's worth noting that we only use LFW for testing.

L2SFO. It is a large-scale synthesized face-with-occlusion dataset built by Yuan et al. [73]. We call it L2SFO in which face images are occluded by six common objects including masks, eyeglasses, sunglasses, cups, scarves, and hands. All the occlusions are located on face images according to segmentation information to augment the reality of this dataset. It contains 991 different identities and more than 73,000 images. We randomly select 891 identities as the training set (about 66,000 images) and the rest as the testing set (about 7000 images).

IJB-C. IARPA Janus Benchmark C is a dataset consisting of video still-frames and photos and used for face recognition benchmark [74]. It contains 117,500 frames from 11,799 videos and 3531 subjects with 31,300 still images. We use the 1:1 protocol for face verification, whose probe and gallery templates are combined using some images and video frames for each subject. Same as the processing procedure of LFW, images in the probe set are occluded and images in the gallery set remain unchanged. We firstly generate clean face images from occluded face images by using our method and other compared methods and then compare the face verification performance. IJB-C is also only used for testing.

4.2. Implementation details

In our experiments, face images are normalized to 256 × 256 and 128 × 128 for high-resolution face completion and face verification, respectively. Following Wu et al. [5], the landmarks in the centers of the eyes and mouth are used for normalizing face images. The occluded face images are generated by MaskTheFace proposed by Anwar and Raychowdhury [75]. We randomly select mask types to occlude face images during training. Some occluded face images are shown in Figs. 4 and 9. For different experimental settings, different datasets are utilized to train our network. For face completion, we train our network on the training sets of CelebA, CelebA-HQ, FFHQ and L2SFO, then testing on their testing sets. As for face verification, we train our network on the training sets of CelebA and Multi-PIE and test on LFW and IJB-C.

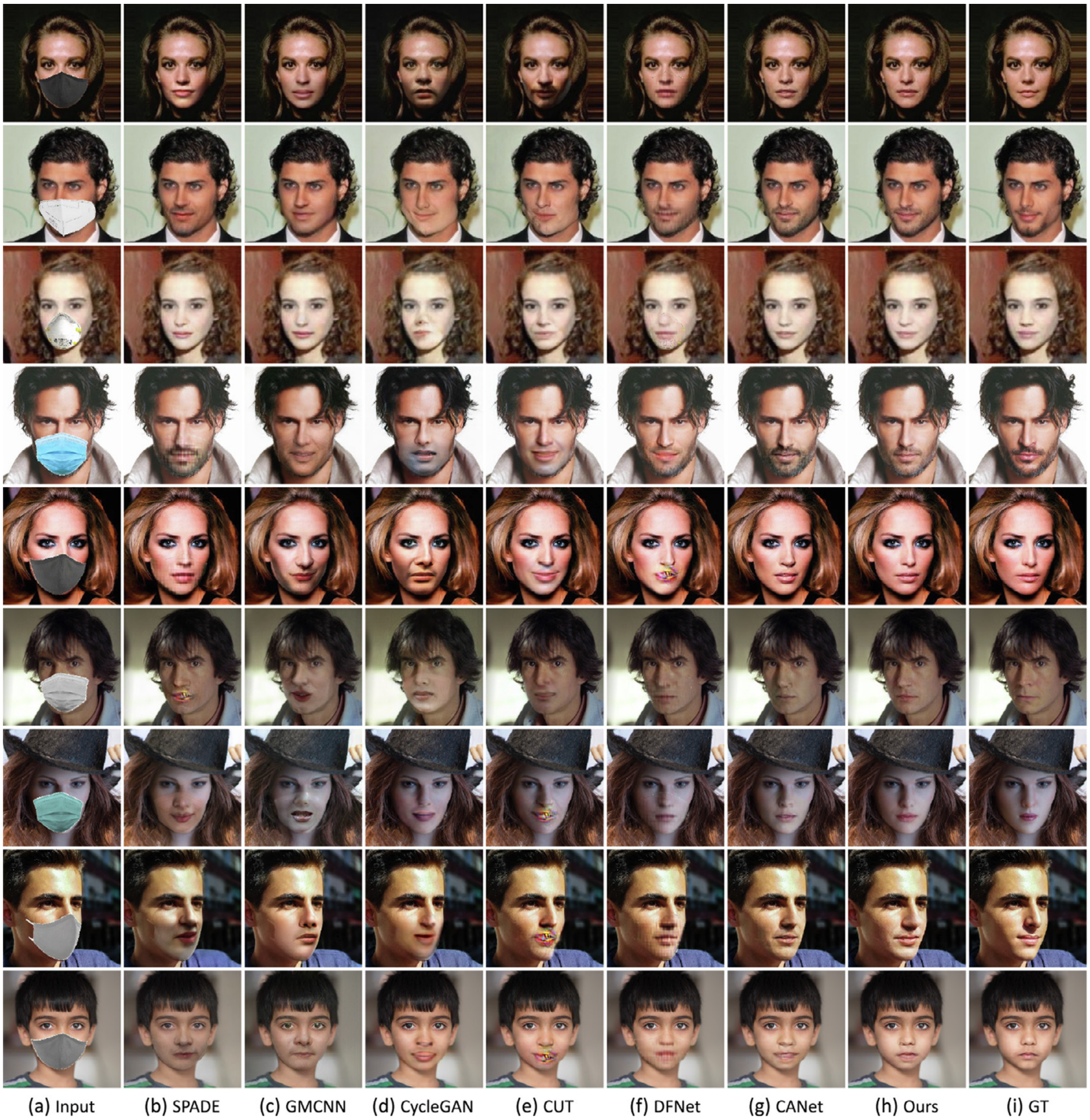

Fig. 4. Qualitative results compared with state-of-the-arts on three datasets. From left to right, (a) are the input images with various kind of masks. (b), (c), (d), (e), (f), (g) and (h) are the results generated by SPADE [76], GMCNN [77], CycleGAN [78], CUT [79], DFNet [20], CANet [27] and ours method respectively. (i) is the ground truth.

Our proposed method can be broken down into two stages. In the first stage, the inference network is trained through contrastive learning until convergence. And in the next stage, the pre-trained encoder and the decoder are jointly trained with the fusion module. We use the SGD optimizer with the learning rate as 0.015 for training the Siamese inference network, and use the Adam optimizer with the learning rate as 10^{-4} for jointly training the encoder and decoder. All the results are reported directly without any additional post-processing. Our proposed method is implemented by the Pytorch framework and trained on four NVIDIA TITAN Xp GPUs (12GB).

4.3. Face completion quantitative results

Peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and Fréchet Inception Distance (FID) are used as evaluation metrics. PSNR and SSIM measure the similarity between the inpainting result and the target image. As for FID, it measures the Wasserstein-2 distance between real and inpainting images through the pre-trained Inception-V3. We select ‘cloth #333333’, ‘KN95’, ‘N95’, ‘surgical blue’, ‘cloth #515151’, ‘surgical’, ‘surgical green’, ‘cloth #dadad9’ and ‘cloth #929292’ masks to occlude the testing images for experiments. These mask images are shown in Fig. 4 from top to bottom.

We conduct quantitative experiments on the testing sets of CelebA, CelebA-HQ and FFHQ occluded by the nine kinds of masks, and report the averaged results. Table 1 shows the performance of our proposed method against other state-of-the-art methods, which consists of two image inpainting methods, GMCNN [77] and DFNet [20], and three image-to-image translation methods: Spade [76], CycleGAN [78] and CUT [79]. In Table 1, we also conduct the experiments to show the improvement of performance compared to our prior conference work [27]. For simplicity, we call it CANet, which can be regarded as a simplified version of our proposed method in this paper without *Dense Correspondence Field Estimation* and the identity preserving loss. We retrain all the compared methods on the training sets of CelebA, CelebA-HQ and FFHQ for the sake of fairness. As shown in Table 1, the proposed method and CANet achieve the best and the second-best quantitative results in three metrics on all the testing sets. The results suggest that the

Table 1
Quantitative comparison on the testing sets of CelebA, CelebA-HQ and FFHQ. †Lower is better. ‡Higher is better.

Dataset	CelebA			CelebA-HQ			FFHQ		
Metric	PSNR‡	SSIM‡	FID†	PSNR‡	SSIM‡	FID†	PSNR‡	SSIM‡	FID†
SPADE [76]	30.92	0.9640	1.8216	27.40	0.9321	30.07	26.27	0.9170	24.45
GMCNN [77]	29.91	0.9563	2.6205	26.10	0.9107	13.07	25.30	0.8963	8.72
CycleGAN [78]	24.47	0.9063	4.9871	21.94	0.8446	13.75	21.13	0.8239	11.26
CUT [79]	24.55	0.9115	5.4059	22.65	0.8690	15.58	21.68	0.8429	12.75
DFNet [20]	32.18	0.9706	4.2948	28.90	0.9494	8.40	29.33	0.9453	11.94
CANet [27]	32.49	0.9731	0.9778	29.89	0.9545	4.23	29.70	0.9501	2.12
Ours	**33.26**	**0.9769**	**0.7981**	**30.67**	**0.9607**	**3.53**	**30.42**	**0.9580**	**1.75**

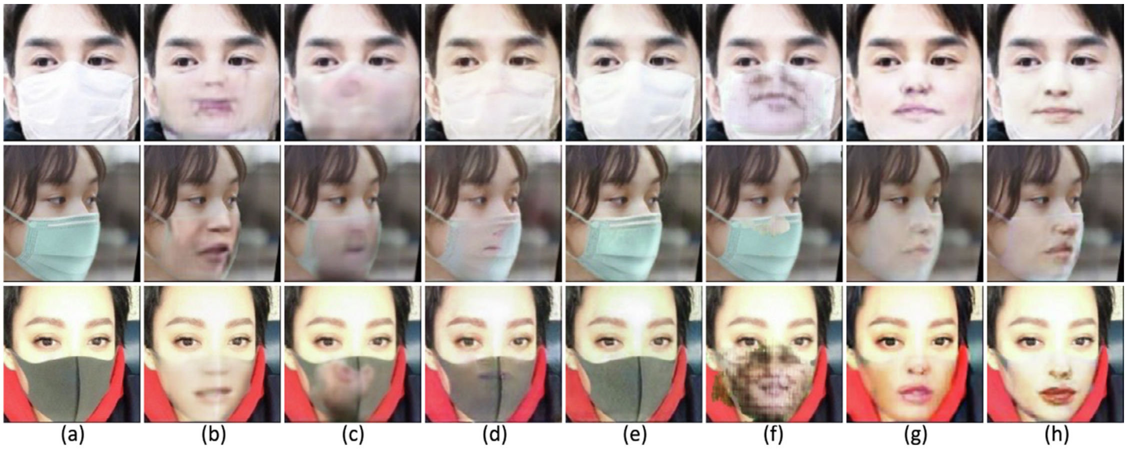

Fig. 5. Face completion results in the wild. (a) is the inputs. (g) is the results generated by our method. (b) - (g) are results produced by SPADE, GMCNN, CycleGAN, CUT, DFNet and CANet, respectively.

proposed method can generate very realistic face images while the compared methods may not work well encountered various kinds of masks. The main reasons for the relatively low performance of the compared methods (excluding CANet) are that 1): face images with various kinds of masks dramatically increase the difficulty of image inpainting, hindering the ability of the representation learning of the encoder; 2): exiting methods take generating realistic images into account but ignore the structural consistency of the generated image. The reason why the performance of our method is higher than CANet may be that *Dense Correspondence Field Estimation* keeps the geometric information of the human face intact during the face completion process.

4.4. Face completion qualitative results

We compare our proposed method with state-of-the-art methods in terms of visual and semantic coherence. We conduct qualitative experiments on the testing sets of three datasets with various kinds of masks. As shown in Fig. 4, we mask the testing images with the nine kinds of masks as described in the last section.

Among all these compared methods, there are severe artifacts in results produced by SPADE, CUT, and DFNet. Thus, the qualities of generated images are far from the requirements. The reason is that various kinds of masks hinder their networks to capture powerful representations. There are no obvious artifacts in face images produced by CycleGAN. But it fails to maintain the geometric information of face images and produce obvious color contrasts. The reason is that CycleGAN endeavors to translate the input to its correspondence non-mask face image and ignores the structural consistency. As for GMCNN, it produces relatively appealing results, but there are significant differences in color at the edges. CANet produces better results in which the facial geometric information is maintained but there are still artifacts, especially in the corners of the mouth. Compared with other methods, our proposed method can generate natural inpainting results with reasonable semantics and richer textures with the help of the self-supervised Siamese inference network, the dense correspondence field, and the DAF module. It demonstrates that our proposed method is superior to the compared methods in terms of consistent structures and colors.

4.5. Face completion in the wild

Furthermore, we also conduct experiments on a real-world masked face dataset (RMFD) [80]. Note that there are no ground truth images in it. Therefore, we directly use our model and the compared models to evaluate on this dataset. As shown in Fig. 5, although there is a huge domain gap between our training sets and the real-world masked face dataset, our method can still generate relatively satisfactory results, which demonstrates the superiority of our proposed method. At the same time, some compared methods can not remove masks effectively, such as (d) and (e) in Fig. 5.

We also provide the corresponding quantitative comparative experiments by using FID, Learned Perceptual Image Similarity (LPIPS) [81], F1-Score and *Realism* in Table 2. LPIPS measures the diversity of images by calculating the similarity in the feature space from the pre-trained AlexNet [82]. F1-Score is the harmonic mean of *recall* and *precision*, where *precision* is calculated by querying whether each generated image is within the estimated manifold of real images and *recall* is calculated by querying whether each real image is within the estimated manifold of generated images [83]. *Realism* is a metric that reflects the distance between the image and the manifold: the closer the image is to the manifold, the higher *Realism* is, and the further the image is from the manifold, the lower Realism is [83]. It clearly demonstrates the superiority of our proposed method in dealing with masked face images in real world.

Table 2
Quantitative comparison on the real world face dataset (RMFD). †Lower is better. ‡Higher is better.

Methods	SPADE	GMCNN	CycleGAN	CUT	DFNet	CANet	Ours
FID †	113.52	150.98	167.73	179.78	173.41	103.34	**98.39**
LPIPS †	0.0827	0.1065	0.1116	0.1303	0.1208	0.0812	**0.0709**
F1-Score ‡	0.026	0.0139	0.0022	0.0034	0.0022	0.0219	**0.0493**
Realism ‡	0.7613	0.7443	0.7089	0.6819	0.6759	0.7723	**0.7883**

Table 3
Face verification results on LFW. 'Masked' means face verification experiments are conducted between the masked probe set and the unchanged gallery set directly.

Model	Metric	Masked	SPADE	GMCNN	CycleGAN	CUT	DFNet	CANet	Ours
ArcFace [84]	AUC	97.51	98.02	97.78	97.65	97.88	97.50	98.09	**98.38**
	FPR=1%	77.71	81.52	79.93	78.28	79.86	78.05	82.12	**83.10**
	FPR=0.1%	44.85	45.56	43.33	44.38	46.43	45.79	50.07	**56.84**
LightCNN [5]	AUC	99.20	99.29	99.30	99.24	99.30	99.20	99.34	**99.49**
	FPR=1%	91.04	92.56	92.63	91.95	92.36	76.87	93.40	**94.41**
	FPR=0.1%	77.13	74.41	77.78	76.06	80.34	64.85	80.34	**82.73**
FaceNet [85]	AUC	98.98	99.08	99.03	99.02	99.03	98.98	99.10	**99.30**
	FPR=1%	85.96	87.51	87.14	86.40	86.30	86.06	87.97	**90.17**
	FPR=0.1%	55.56	57.07	53.10	54.31	56.06	55.72	56.03	**57.85**

Table 4
Quantitative comparison on the L2SFO dataset. †Lower is better. ‡Higher is better.

Methods	SPADE	GMCNN	CycleGAN	CUT	DFNet	CANet	Ours
PSNR ‡	22.6	22.57	20.11	20.2	22.73	23.3	**23.69**
SSIM ‡	0.8775	0.8785	0.8297	0.8292	0.8805	0.8975	**0.9007**
FID †	40.59	44.31	64.09	62.12	33.04	31.6	**29.14**

Table 5
Face verification results on IJB-C. 'Masked' means face verification experiments are conducted between the masked probe set and the unchanged gallery set directly.

Model	Metric	Masked	SPADE	GMCNN	CycleGAN	CUT	DFNet	CANet	Ours
ArcFace	AUC	97.46	97.60	97.68	97.56	97.57	97.66	97.77	**98.03**
	FPR=1%	80.07	80.98	81.60	81.18	81.14	81.61	82.23	**84.38**
	FPR=0.1%	60.53	61.73	62.49	62.41	61.93	62.86	63.33	**67.21**
LightCNN	AUC	99.13	99.16	99.15	99.14	99.14	99.14	99.14	**99.20**
	FPR=1%	93.50	93.90	93.70	93.68	93.56	93.68	93.88	**94.62**
	FPR=0.1%	83.99	84.97	84.50	84.47	84.22	84.61	85.31	**87.20**
FaceNet	AUC	99.15	99.18	99.15	99.17	99.16	99.18	99.22	**99.27**
	FPR=1%	91.76	92.11	91.67	91.73	91.82	91.86	92.40	**92.94**
	FPR=0.1%	78.47	79.15	78.43	78.72	78.59	79.14	79.72	**80.88**

4.6. Face completion on free-Form occlusions

In the above three sections, we mainly conduct quantitative and qualitative experiments on face images with masks. In order to demonstrate the effectiveness of our method, we conduct experiments on the L2SFO dataset [73] in which face images are occluded by six common objects, i.e, masks, eyeglasses, sunglasses, cups, scarves, and hands. We conduct quantitative experiments on the testing set of L2SFO, and report the averaged results. we also retrain all the compared methods on the training sets of L2SFO for the sake of fairness. Table 4 shows the performance of our proposed method against other compared methods. Our method outperforms all the other compared methods in three metrics on the testing sets as shown in this table. The results suggest that the proposed method can still extend to other kinds of occlusions.

We also compare our proposed method with the state-of-the-art methods in terms of the visual quality on the testing set of L2SFO. As shown in Fig. 6, we find that SPADE and GMCNN can remove occlusions, but there are serious artifacts in the generated images. CycleGAN and CUT fail to remove occlusions in some cases. Because they adopt unsupervised learning and hardly handle face images with complex occlusions. DFNet and CANet achieve relatively high-quality results. However, there are still artifacts in the generated face images produced by them. Different from all the compared methods, the proposed method can generate photo-realistic face images.

4.7. Face verification results

In order to quantitatively evaluate the feasibility of our method for face verification, we compare the results of our method and the compared methods on LFW and IJB-C following the testing protocol as described in Sec 4.1. Face verification experiments are conducted between the recovered probe set and the unchanged gallery set. Three publicly released face recognition models are tested: the LightCNN [5], ArcFace [84] and FaceNet [85]. We use the area under the ROC curve (AUC), true positive rates at 1% and 0.1% (TPR@FPR=1%, TPR@FPR=0.1%) as the evaluation metrics in the experiments. The results are reported in Tables 3 and 5.

We use the masked probe set as a baseline to demonstrate the influences of face completion on face verification. From Table 3, we can see that our method brings dramatic improvement to face verification. Because our method can keep geometric information intact and generate face images with consistent structures and colors. Compared with the baseline, our method can achieve an in-

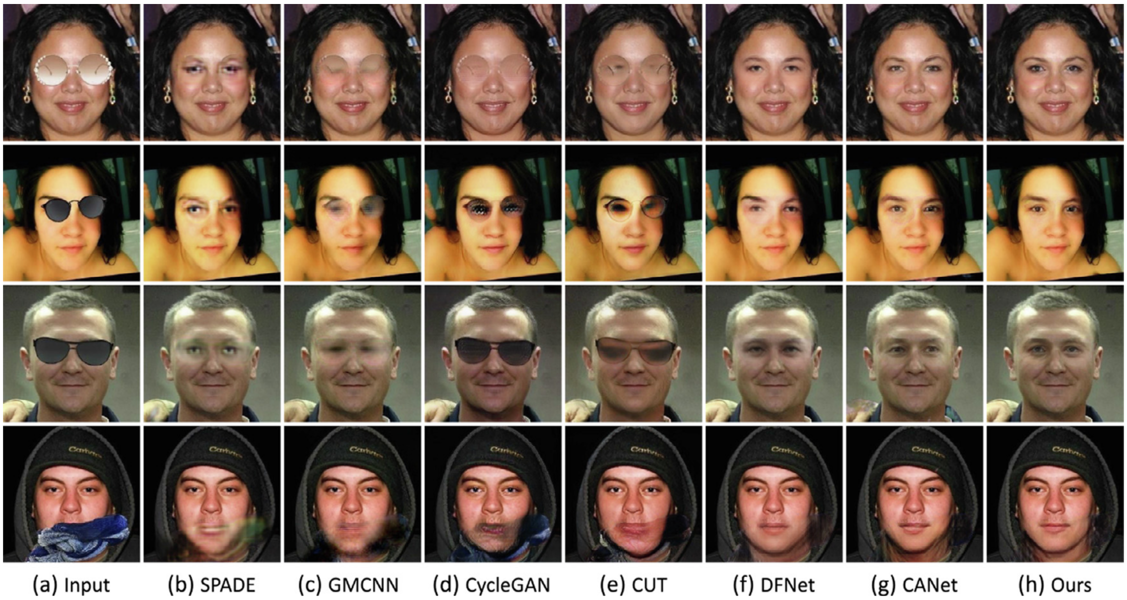

Fig. 6. Face completion results on the L2SFO dataset. From left to right, (a) are the input images. (b), (c), (d), (e), (f), (g) and (h) are the results generated by SPADE, GMCNN, CycleGAN, CUT, DFNet, CANet and ours method respectively..

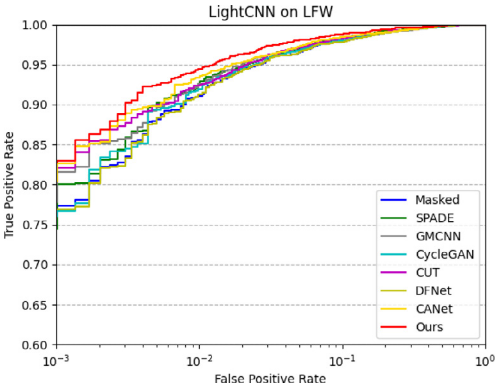

Fig. 7. The ROC curves on the LFW dataset using LightCNN.

crease of more than 10% in TPR@FPR=0.1% on LFW and an increase of 6.68% in TPR@FPR=0.1% on IJB-C, which demonstrates that our proposed method can ameliorate the negative impact of masks. Similar to our method, the compared methods endeavor to recover face images. However, we find that the face verification performances of some compared methods decrease actually, especially in TPR@FPR=0.1%. For instance, the performance of CycleGAN drops from 77.13% to 76.06% on LFW, a drop of about 1% when taking the metric TPR@FAR=1% and using LightCNN as the face feature extractor. From Table 5, we can also see that the compared methods do not show obvious advantages over the baseline ('Masked') on IJB-C. For example, the performance of CUT is 91.82%, a very limited improvement of 0.006% over the baseline when taking the metric TPR@FAR=1% and using FaceNet as the face feature extractor. For the poor performances of compared methods on LFW and IJB-C, the reason may lie in two aspects. The first reason is that the compared methods can not generate high-quality face images. The other reason is that they can not recover discriminative information of a face image due to the great negative effects of masks. We also present the ROC curves on LFW in Fig. 7. It is obvious that our method outperforms all the compared methods.

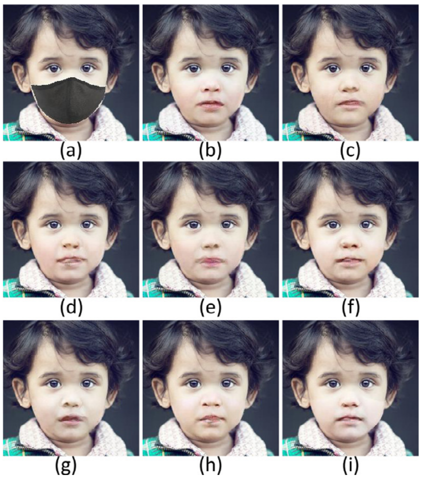

Fig. 8. Images produced by the variants of our proposed method. (a) is the input with the 'cloth #33333' mask. (i) is the result generated by the full model. (b)-(h) are results generated by the variant models according to Table 7.

Table 6
The inference time (seconds) on GPU and CPU.

Inference Time	SPADE	GMCNN	CycleGAN	CUT	DFNet	Ours
GPU	0.023	0.023	0.025	0.006	0.019	0.017
CPU	1.505	1.474	4.132	1.075	0.188	0.248

4.8. Time complexity

We conduct the time complexity experiments on a single GPU (TITAN Xp) and CPU, respectively. To evaluate the inference time for different methods, we randomly sample 1000 testing images and run forward one time for each image. Then we report the mean inference time for one image. As shown in Table 6, our proposed method achieves a pleasing time performance compared with the other methods. It runs the second fast on a single TITAN Xp GPU. The fastest method is CUT on GPU. Because the number of parameters of CUT is only about a quarter of our method. However, as can be seen from Table 1 and Fig. 4, our method outper-

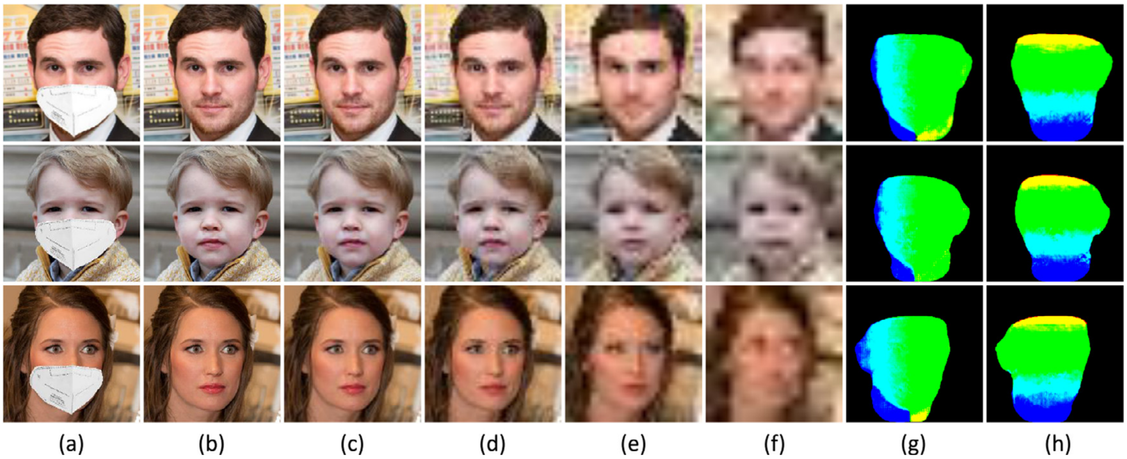

Fig. 9. Images produced by the multi-scale decoder. (a) is the inputs with a 'KN95' mask. (b) is the final inpainting results. (c), (d), (e) and (f) are outputs at multi-scale. (g) and (h) are the estimated U and V maps, respectively.

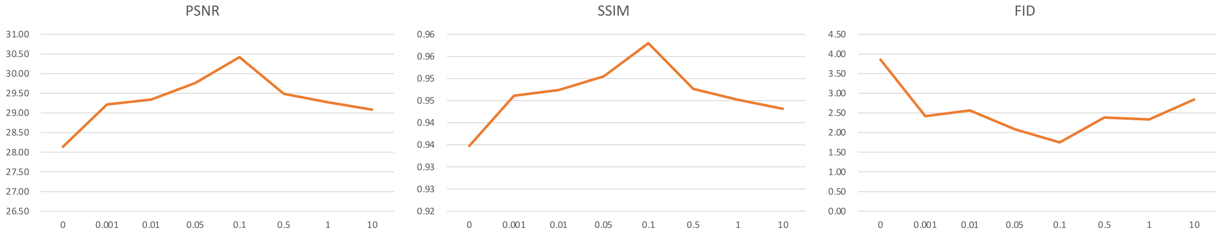

Fig. 10. Model performance affected by the weight of the UV loss on the FFHQ dataset..

Table 7
Ablation study experiments on the testing set of CelebA. †Lower is better. ‡Higher is better. CL means Contrastive Learning.

CL	✗	√	✗	✗	√	√	✗	√
DAF	✗	✗	√	✗	√	✗	√	√
UV map	✗	✗	✗	√	✗	√	√	√
PSNR ‡	29.71	31.04	31.82	31.11	32.02	32.20	32.40	**32.82**
SSIM ‡	0.9568	0.9664	0.9674	0.9663	0.9702	0.9708	0.9723	**0.9755**
FID †	1.9657	1.4729	1.3899	1.5059	1.259	1.3806	0.9872	**0.9040**

forms CUT with a large margin. When running on CPU, our proposed method is faster than SPADE, GMCNN, CycleGAN and CUT and achieves the comparable performance against DFNet.

4.9. Ablation study

We investigate the effectiveness of different components of the proposed method on the testing set of CelebA. We train several variants of the proposed method: remove the self-supervised Siamese inference network (denote as contrastive learning), the DAF module, and/or the dense correspondence estimation (denoted as UV map). As shown in Table 7, it clearly demonstrates that the self-supervised Siamese inference network, the DAF module, and the dense correspondence field estimation play important roles in determining the performance. As shown in Fig. 8, the uncompleted models usually generate images with obvious artifacts, especially in boundaries while our full model can suppress color discrepancy and artifacts in boundaries and produce realistic inpainting results.

The multi-scale decoder can progressively refine the inpainting results at each scale. We also conduct experiments on the testing set of FFHQ. Then we visualize the images predicted by the decoder at several scales. As shown in Fig. 9, it demonstrates that this multi-scale architecture is beneficial for decoding learned representations into generated images layer by layer.

We conduct sufficient experiments on the FFHQ dataset to explore the performance variation of our model affected by the weight of the UV loss function. We plot some figures according to the experimental results (Fig. 10). The horizontal axis represents the weight of the UV loss function. We use eight different weights to design the experiment, i.e, 0, 0.001, 0.01 0.05, 0.1, 0.5, 1 and 10. From Fig. 10, we can see that PSNR gradually increases with the increase of weight, reaches the maximum value when weight is equal to 0.1, and then drops sharply. The variation of SSIM is roughly the same as that of PSNR. The value of FID decreases dramatically from about 4 at the weight of 0 to around 2.5 at the weight of 0.001 and reaches the bottom (about 1.7) at the weight of 0.1. From these experiments, we can see that the UV loss (or *Dense Correspondence Field Estimation*) plays an important role in determining the performance since it can keep the geometric information of the human face intact during the face completion process.

5. Conclusion

In this paper, we propose a novel two-stage paradigm image inpainting method to generate smoother results with reasonable semantics and richer textures. Specifically, the proposed method boosts the ability of the representation learning of the inference network by using contrastive learning. For keeping the geometric information of the input face image intact, we introduce a dense correspondence field that binds the 2D and 3D surface spaces into our network. We further design a novel dual attention fusion module, which can be embedded into decoder layers in a plug-and-play way. Extensive experiments show the superiority of our proposed method in generating smoother, more coherent, and fine-detailed results, and demonstrate our method can greatly improve the performance of face verification.

Declaration of Competing Interest

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We further confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

Acknowledgements

This work is partially funded by National Natural Science Foundatio of China (Grant No. 62006228).

References

[1] M. Bertalmio, G. Sapiro, V. Caselles, C. Ballester, Image inpainting, in: International Conference on Computer Graphics and Interactive Techniques, 2000, pp. 417–424.

[2] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Free-form image inpainting with gated convolution, in: IEEE International Conference on Computer Vision, 2019, pp. 4471–4480.

[3] C. Barnes, E. Shechtman, A. Finkelstein, D.B. Goldman, PatchMatch: a randomized correspondence algorithm for structural image editing, ACM Trans. Graph. 28 (3) (2009) 24–33.

[5] X. Wu, R. He, Z. Sun, T. Tan, A light CNN for deep face representation with noisy labels, IEEE Trans. Inf. Forensics Secur. 13 (11) (2018) 2884–2896.

[8] A.A. Efros, W.T. Freeman, Image quilting for texture synthesis and transfer, in: International Conference on Computer Graphics and Interactive Techniques, 2001, pp. 341–346.

[9] Z. Xu, J. Sun, Image inpainting by patch propagation using patch sparsity, IEEE Trans. Image Process. 19 (5) (2010) 1153–1165.

[12] J. Cao, Y. Hu, H. Zhang, R. He, Z. Sun, Learning a high fidelity pose invariant model for high-resolution face frontalization, in: Advances in Neural Information Processing Systems, 2018, pp. 2867–2877.

[13] H. Huang, R. He, Z. Sun, T. Tan, Wavelet domain generative adversarial network for multi-scale face hallucination, Int. J. Comput. Vis. 127 (6–7) (2019) 763–784.

[14] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, E. Ding, Image inpainting with learnable bidirectional attention maps, in: IEEE International Conference on Computer Vision, 2019, pp. 8858–8867.

[15] G. Liu, F.A. Reda, K.J. Shih, T.-C. Wang, A. Tao, B. Catanzaro, Image inpainting for irregular holes using partial convolutions, in: European Conference on Computer Vision, 2018, pp. 85–100.

[16] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.

[17] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, IEEE, 2006, pp. 1735–1742.

[18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514.

[19] L. Chen, Topological structure in visual perception, Science 218 (4573) (1982) 699–700.

[20] X. Hong, P. Xiong, R. Ji, H. Fan, Deep fusion network for image completion, in: ACM International Conference on Multimedia, 2019, pp. 2033–2042.

[21] P. Pérez, M. Gangnet, A. Blake, Poisson image editing, in: ACM SIGGRAPH 2003 Papers, 2003, pp. 313–318.

[22] S. Zhang, R. He, Z. Sun, T. Tan, DeMeshNet: blind face inpainting for deep meshface verification, IEEE Trans. Inf. Forensics Secur. 13 (3) (2017) 637–647.

[23] Z. Li, Y. Hu, R. He, Z. Sun, Learning disentangling and fusing networks for face completion under structured occlusions, Pattern Recognit. 99 (2020) 107073.

[24] J. Cai, H. Han, J. Cui, J. Chen, L. Liu, S.K. Zhou, Semi-supervised natural face de-occlusion, IEEE Trans. Inf. Forensics Secur. 16 (2020) 1044–1057.

[25] R. Alp Guler, G. Trigeorgis, E. Antonakos, P. Snape, S. Zafeiriou, I. Kokkinos, DenseReg: fully convolutional dense shape regression in-the-wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6799–6808.

[26] R. Alp Güler, N. Neverova, I. Kokkinos, DensePose: dense human pose estimation in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.

[27] X. Ma, X. Zhou, H. Huang, Z. Chai, X. Wei, R. He, Free-form image inpainting via contrastive attention network, in: International Conference on Pattern Recognition, 2020.

[28] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410.

[29] A. Criminisi, P. Pérez, K. Toyama, Region filling and object removal by exemplar-based image inpainting, IEEE Trans. Image Process. 13 (9) (2004) 1200–1212.

[30] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: feature learning by inpainting, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.

[31] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.

[36] S. Iizuka, E. Simo-Serra, H. Ishikawa, Globally and locally consistent image completion, ACM Trans. Graph. 36 (4) (2017) 1–14.

[37] K. Nazeri, E. Ng, T. Joseph, F.Z. Qureshi, M. Ebrahimi, EdgeConnect: generative image inpainting with adversarial edge learning, arXiv preprint arXiv:1901.00212(2019).

[38] Y. Ren, X. Yu, R. Zhang, T.H. Li, S. Liu, G. Li, StructureFlow: image inpainting via structure-aware appearance flow, in: IEEE International Conference on Computer Vision, 2019, pp. 181–190.

[39] Y. Zeng, Z. Lin, J. Yang, J. Zhang, E. Shechtman, H. Lu, High-resolution image inpainting with iterative confidence feedback and guided upsampling, in: European Conference on Computer Vision, 2020.

[40] J. Cai, H. Han, S. Shan, X. Chen, FCSR-GAN: joint face completion and super-resolution via multi-task learning, IEEE Trans. Biom. Behav.Identity Sci. 2 (2) (2019) 109–121.

[41] T. Zhou, C. Ding, S. Lin, X. Wang, D. Tao, Learning oracle attention for high–fidelity face completion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 7680–7689.

[42] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, C.C. Loy, Self-supervised scene de-occlusion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[43] T. Chen, S. Kornblith, M. Norouzi, G.E. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, 2020.

[44] S. Becker, G.E. Hinton, Self-organizing neural network that discovers surfaces in random-dot stereograms, Nature 355 (6356) (1992) 161–163.

[45] A. Dosovitskiy, J.T. Springenberg, M. Riedmiller, T. Brox, Discriminative unsupervised feature learning with convolutional neural networks, Adv. Neural Inf. Process. Syst. 27 (2014) 766–774.

[46] Z. Wu, Y. Xiong, S.X. Yu, D. Lin, Unsupervised feature learning via non-parametric instance discrimination, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3733–3742.

[47] C. Zhuang, A.L. Zhai, D. Yamins, Local aggregation for unsupervised learning of visual embeddings, in: IEEE International Conference on Computer Vision, 2019, pp. 6002–6012.

[48] S.K. Mustikovela, V. Jampani, S.D. Mello, S. Liu, U. Iqbal, C. Rother, J. Kautz, Self-supervised viewpoint learning from image collections, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3971–3981.

[49] X. Zhan, X. Pan, B. Dai, Z. Liu, D. Lin, C.C. Loy, Self-supervised scene de-occlusion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 3784–3792.

[50] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.

[53] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.

[54] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, in: International Conference on Computer Graphics and Interactive Techniques, 1999, pp. 187–194.

[55] Y. Li, H. Huang, J. Cao, R. He, T. Tan, Disentangled representation learning of makeup portraits in the wild, Int. J. Comput. Vis. (2019) 1–19.

[56] X. Tu, J. Zhao, Z. Jiang, Y. Luo, M. Xie, Y. Zhao, L. He, Z. Ma, J. Feng, Joint 3d face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning, arXiv preprint arXiv:1903.09359(2019).

[57] J. Roth, Y. Tong, X. Liu, Unconstrained 3D face reconstruction, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2606–2615.

[58] X. Tu, J. Zhao, M. Xie, Z. Jiang, A. Balamurugan, Y. Luo, Y. Zhao, L. He, Z. Ma, J. Feng, 3D face reconstruction from a single image assisted by 2D face images in the wild, IEEE Trans. Multimedia (2020).

[59] X. Wang, A. Gupta, Unsupervised learning of visual representations using videos, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2794–2802.

[60] P. Bachman, R.D. Hjelm, W. Buchwalter, Learning representations by maximizing mutual information across views, in: Advances in Neural Information Processing Systems, 2019, pp. 15535–15545.

[61] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, Z. Liu, Dynamic convolution: Attention over convolution kernels, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 11030–11039.

[62] T. Karras, T. Aila, S. Laine, J. Lehtinen, Progressive growing of GANs for improved quality, stability, and variation, in: International Conference on Learning Representations, 2018.

[64] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image super-resolution using very deep residual channel attention networks, in: European Conference on Computer Vision, 2018, pp. 286–301.

[65] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, A 3D face model for pose and illumination invariant face recognition, in: IEEE International Conference on Advanced Video and Signal Based Surveillance, Ieee, 2009, pp. 296–301.

[66] X. Zhu, Z. Lei, X. Liu, H. Shi, S.Z. Li, Face alignment across large poses: a 3D solution, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 146–155.

[67] J. Booth, S. Zafeiriou, Optimal uv spaces for facial morphable model construction, in: International Conference on Image Processing, IEEE, 2014, pp. 4672–4676.

[68] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.

[69] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of International Conference on Computer Vision (ICCV), 2015.

[70] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (5) (2010) 807–813.

[71] R. Huang, S. Zhang, T. Li, R. He, Beyond face rotation: global and local perception GAN for photorealistic and identity preserving frontal view synthesis, in: IEEE International Conference on Computer Vision, 2017, pp. 2439–2448.

[72] G.B. Huang, M. Mattar, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database forstudying face recognition in unconstrained environments, 2008.

[73] X. Yuan, I.K. Park, Face de-occlusion using 3D morphable model and generative adversarial network, in: IEEE International Conference on Computer Vision, 2019, pp. 10062–10071.

[74] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J.C. Adams, T. Miller, N.D. Kalka, A.K. Jain, J.A. Duncan, K.E. Allen, J. Cheney, P. Grother, IARPA janus benchmark-b face dataset, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 592–600.

[75] A. Anwar, A. Raychowdhury, Masked face recognition for secure authentication, arXiv preprint arXiv:2008.11104(2020).

[76] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.

[77] Y. Wang, X. Tao, X. Qi, X. Shen, J. Jia, Image inpainting via generative multi-column convolutional neural networks, in: Advances in Neural Information Processing Systems, 2018, pp. 331–340.

[78] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.

[79] T. Park, A.A. Efros, R. Zhang, J.-Y. Zhu, Contrastive learning for unpaired image–to-image translation, in: European Conference on Computer Vision, Springer, 2020, pp. 319–345.

[80] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, et al., Masked face recognition dataset and application, arXiv preprint arXiv:2003.09093(2020).

[81] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.

[82] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, Adv. Neural Inf. Process Syst. 25 (2012) 1097–1105.

[83] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, T. Aila, Improved precision and recall metric for assessing generative models, Adv. Neural Inf. Process. Syst. (2019).

[84] J. Deng, J. Guo, N. Xue, S. Zafeiriou, ArcFace: additive angular margin loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4690–4699.

[85] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[86] Q. Wang, H. Fan, G. Sun, Y. Cong, Y. Tang, Laplacian pyramid adversarial network for face completion, Pattern Recognit. 88 (2019) 493–505.

[87] M. Luo, X. Ma, Z. Li, J. Cao, R. He, Partial nir-vis heterogeneous face recognition with automatic saliency search, IEEE Trans. Inf. Forensics Secur. (2021).

[88] Z. Pei, M. Jin, Y. Zhang, M. Ma, Y.-H. Yang, All-in-focus synthetic aperture imaging using generative adversarial network-based semantic inpainting, Pattern Recognit. 111 (2021).

[89] D. Ding, S. Ram, J. J. Rodríguez, Image inpainting using nonlocal texture matching and nonlinear filtering, IEEE Trans. Image Process. 28 (4) (2018) 1705–1719.

[90] A. M. Obeso, J. Benois-Pineau, M. S. G. Vázquez, A. Á. R. Acosta, Visual vs internal attention mechanisms in deep neural networks for image classification and object detection, Pattern Recognit. 123 (2021).

[91] R. Chen, H. Zhang, J. Liu, Multi-attention augmented network for single image superresolution, Pattern Recognit. 122 (2022).

[92] Y. Zeng, Y. Gong, J. Zhang, Feature learning and patch matching for diverse image inpainting, Pattern Recognit. 119 (2021).

[93] D. Ding, S. Ram, J. J. Rodriguez, Perceptually aware image inpainting, Pattern Recognit. 83 (2018) 174–184.

[94] D. Kim, S. Woo, J.-Y. Lee, I. S. Kweon, Recurrent temporal aggregation framework for deep video inpainting, IEEE Trans. Pattern Anal. Mach. Intell. 42 (5) (2019) 1038–1052.

[95] W. He, Q. Yao, C. Li, N. Yokoya, Q. Zhao, H. Zhang, L. Zhang, Non-local meets global: An integrated paradigm for hyperspectral image restoration, IEEE Trans. Pattern Anal. Mach. Intell. (2020).

[96] N. Wang, S. Ma, J. Li, Y. Zhang, L. Zhang, Multistage attention network for image inpainting, Pattern Recognit. 106 (2020).