SwapInpaint: Identity-Specific Face Inpainting With Identity Swapping

Honglei Li, Wenmin Wang, *Member, IEEE*, Cheng Yu, and Shixiong Zhang

***Abstract*— As face editing scenarios have become popular, the face inpainting technique has become a hot topic. Although some existing methods can inpaint faces with preserved identity information, they fail to solve a more flexible inpainting problem that fills the "holes" with identity-specific content from other faces. In this work, we propose a disentangle and subject-agnostic framework that affects both full and partial-face inpainting with the guidance of a reference face image. The framework consists of an identity encoding module, a content inference module and a generative module. The identity encoding module extracts the identity embedding from the reference image, the content inference module learns to predict the content image, and the generative module integrates the content image and the reference identity embedding to generate the identity-specific inpainted result. To minimize the structure and style gap between the incomplete image and inpainted image, we use a double attribute loss to the generative module and a postprocess of blending operation to the swapped result. We compare our method with state-of-the-art works and demonstrate that our method achieves higher identity similarity and better structural correctness.**

***Index Terms*— Identity-specific inpainting, partial-face inpainting, face swapping.**

I. Introduction

In RECENT years, face inpainting has become an important technique and is widely used in face editing tasks, including eye-inpainting [1], blemish removal [2], face appearance [3], attribute transferal [4], [5], and photo restoration [6], [7]. However, humans are extremely sensitive to identify inconsistency, especially when the faces in the photo are well known [8], and identity-preserved photo inpainting technologies have universal practical significance. Furthermore, more flexible face editing by replacing the partial-face region (e.g., eye or mouse region) with corresponding content from other people is desirable. In this paper, we focus on solving this problem by taking a reference face image as guidance and producing inpainted results with identity-specific content, which we refer to as **identity-specific face inpainting**, as shown in Figure 1.

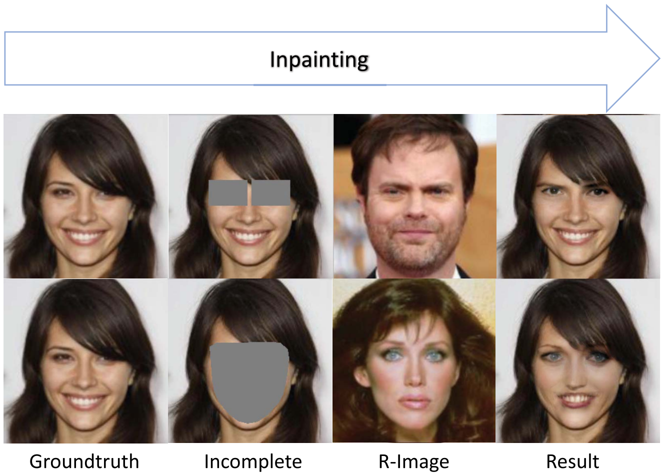

Fig. 1. Our algorithm takes the reference image as guidance and produces inpainted results with identity-specific content.

Current generic inpainting methods [9]–[12] treat inpainting as a conditional generative problem and fill "holes" by combining autoencoder [13] or variational autoencoder (VAE [14]) networks with a GAN [15] discriminator. These methods transfer the incomplete images to the latent domain by an encoder and then generate semantic inpainting results according to the latent embeddings, as shown in Figure 2(a). Autoencoder-based methods always produce deterministic results, and VAE-based methods (e.g., PICNet [12]) are used to produce results with more diversity. Some face inpainting methods [16]–[19] improve the face structure by adding facial guidance or designing facial concerning loss. As there is no identity information offered, these methods rarely generate identity-specific content but generate the content most likely globally consistent. Early identity-preserving method ExGAN [1] solves the eye-inpainting task by using an exemplar image or a reference code to produce personalized inpainting results, as shown in Figure 2(b). Zhao *et al.* [20] solved the identity-preserving face completion problem by using identity features to supervise the training process, as shown in Figure 2(c). However, it is trained to preserve the identity information of the full-face region. When filling partial "holes" (e.g., eye or mouse region) with the guidance of other people, it becomes difficult to restrict the result with identity features that are extracted from a full face.

Manuscript received November 2, 2021; accepted November 16, 2021. Date of publication November 23, 2021; date of current version July 5, 2022. This work was supported by the Science and Technology Development Fund (FDCT) of Macau under Grant 0016/2019/A1. This article was recommended by Associate Editor Z. Tang. *(Corresponding author: Wenmin Wang.)*

Honglei Li is with the International Institute of Next Generation Internet, Macau University of Science and Technology, Macau, China, and also with the Chongqing College of Electronic Engineering, Chongqing, China (e-mail: l_honglei@foxmail.com).

Wenmin Wang, Cheng Yu, and Shixiong Zhang are with the International Institute of Next Generation Internet, Macau University of Science and Technology, Macau, China (e-mail: wmwang@must.edu.mo).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3130196.

Digital Object Identifier 10.1109/TCSVT.2021.3130196

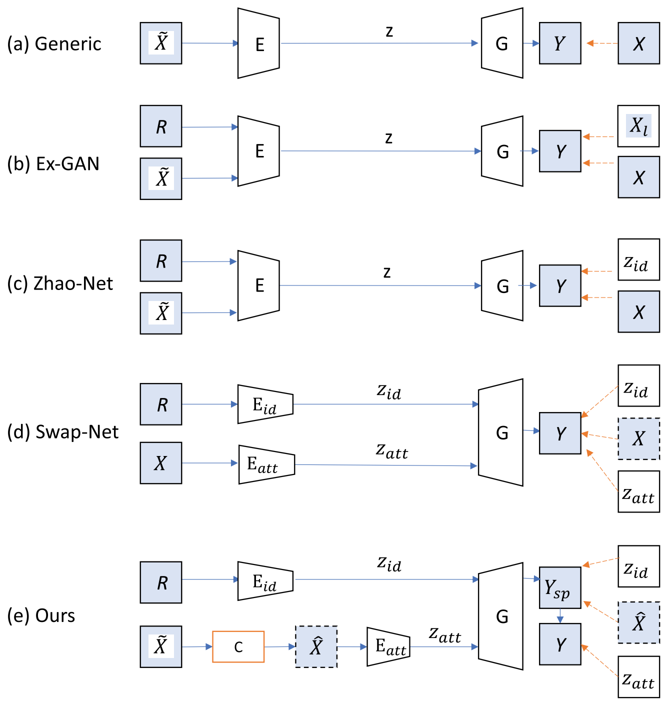

Fig. 2. Illustration of the generic inpainting network structures, ExGAN [1], Zhao-Net [20], Swap-Net [24] and the proposed SwapInpaint. $\tilde{X}$ and R are incomplete images and reference images. X denotes groud truths. E, G, and C are the encoder, generator and content inferrer, respectively. z, z_{id}, and z_{att} are the generic, identity and attribute latent codes, which are generated from E, E_{id}, and E_{att}, respectively. $\hat{X}$ is the content-inferred result, which fills the holes filled by the C network. Y_{sp} and Y are the swapped outputs and cropped outputs from G.

Face swapping can be treated as a full-face inpainting technique that fills the target face "holes" with identity-specific content that is the same as the reference face. Some subject-specific methods (e.g., DeepFake [21] and fast face-swap [22]) train a new model for any new face pairs, which is not practical in our problem. Recent subject-agnostic methods [23]–[27], which train only one model for swapping any face pairs, easily implement full-face inpainting with the help of facial masks. For example, Gu-Net [27] learns a face synthesis model to inpaint a target face by the use of component embeddings from the source image and a manually edited facial mask. However, it is difficult to draw a precise facial mask when some target pixels are missing. The more recent FaceShifter [24] generates high-fidelity results by combining content attribute embeddings from the target face and identity embedding from the reference face, and its network structure is similar to Figure 2(d). However, when changing the target input into an incomplete image, the attribute encoder fails to extract reasonable content attributes and generate inconsistent structure results, as shown in Figure 6.

Inspired by PICNet [12] and FaceShifter [24], we propose a new face inpainting network named **SwapInpaint**, which uses a content inferrer to supplement attribute information, as shown in Figure 2(e). Our method disentangles the content inference and identity swapping, conducting an inpainting task into three stages: 1) use a content inference module to fill the "holes" with structure-correct content, 2) swap the identity with the guidance of identity code extracted from a reference face image through a subject-agnostic face-swap module, and 3) adopt a blending operation to eliminate style gaps. As the swap is a whole image reconstruction process, the results may be biased in style and posture. However, there is no good method for making the result consistent in posture while keeping the style consistent. In our work, we assume that the inpainted result has the same posture as the swapped result and address the posture difference by using double attribute loss in the generative model to supervise the posture consistency.

The framework we proposed in this paper affects both full and partial-face inpainting with the guidance of the reference face image, which can be applied to flexible face inpainting and face swapping. Our main contributions in this work are as follows:

1) We propose a disentangle and subject-agnostic identity-specific inpainting method named SwapInpaint, which combines content inference and identity swap to solve the challenging problem of identity-specific face inpainting, especially partial-face inpainting.
2) We design a new network structure according to the SwapInpaint method to improve the ability to handle inpainting tasks.
3) We conduct experiments to demonstrate that our work can fill face "holes" with the content of better structure correctness and identity similarity, as examples shown in Figure 4.

II. Related Work

A. Image Inpainting

Image inpainting is the technique to fill the "holes" with plausible content and has been broadly applied in visual content editing. Early image inpainting works [28]–[34] are driven by data similarity, focus on texture synthesis [30], [35] or patch matching techniques [29], by assuming that missing regions should be filled with similar textures or patches from the known region. Although these methods are successfully used in low-level texture recovery tasks, they fail to capture the global structure and generate semantic content for large holes.

In recent years, learning-based image inpainting methods [9], [36]–[40] have become mainstream, and these methods have been proposed to generate semantic content by learning the underlying distribution from a large quantity of training data. Learning-based approaches can be grouped into two schools: single-stage and multistage. The first school trains the model through a single encoder-decoder-discriminator structured network, which uses an encoder to learn the latent distribution of the incomplete images, a decoder to reconstruct missing content from the learned distribution, and a discriminator to increase fidelity. The ContextEncoder [38] is the earliest one-stage inpainting model with the ability to fill 64×64 -sized center holes. The following single-stage methods improve models by combining global and local adversarial loss [9] or designing particular convolutions to support irregular hole filling [36], [37], [39], [40]. Their common drawback is that they often create blurry content, especially for large

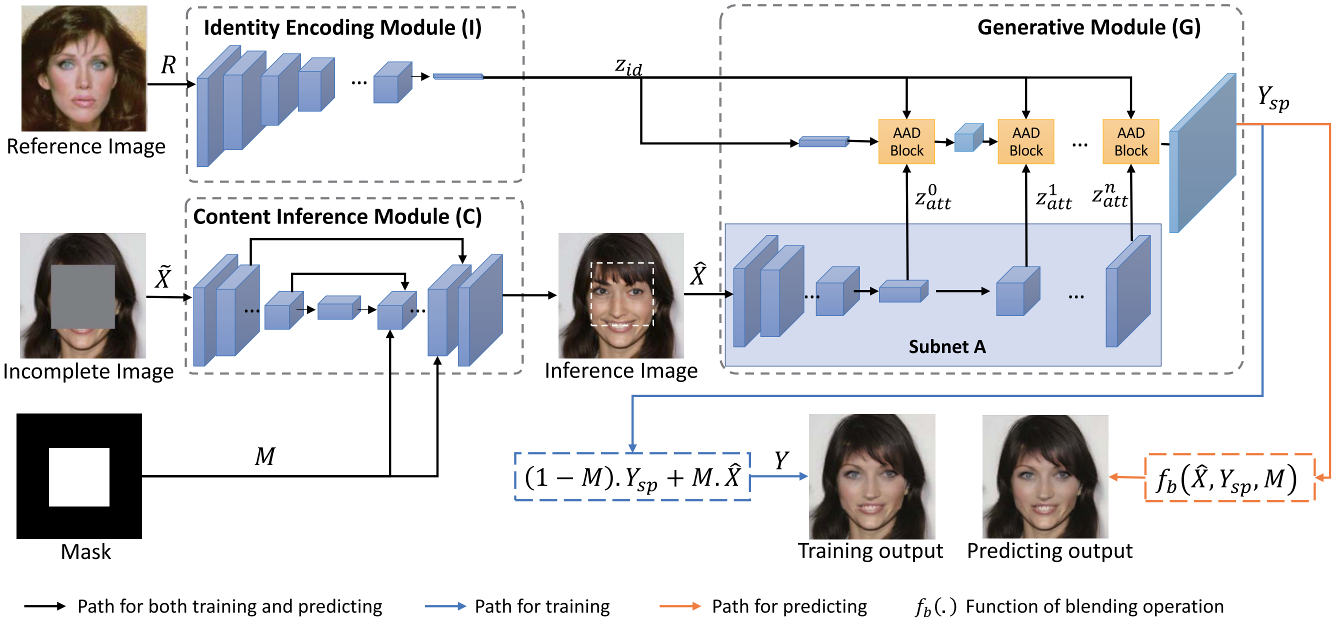

Fig. 3. The overall architecture of our SwapInpaint network consists of an identity encoding module (I), a content inference module (C), and a generative module (G). Module C is trained to infer the missing content of incomplete image $\tilde{X}$ to produce inference image $\hat{X}$. The generative module G integrates the identity embedding z_{id} and inference image $\hat{X}$ to generate swapped face Y_{sp} and the training result Y in the training process. In the prediction process, a blending operation $f(.)$ is used to remove the style discontinuities.

holes. More recently, usage of attention [10] and VAEs [14] greatly improved the global structure consistency and fidelity. Methods, e.g., [12], [41] design a contextual attention layer to capture more semantically generative features. VAEs are exploited by some methods [4], [12], [42] to produce diverse outputs. The second school divides the learning process into two or more stages, which can be pretrained models or networks designed separately to satisfy special training strategies. Iterative inference methods [36], [43] utilize a pretrained generative model to infer a suitable latent prior for "hole" completion. Some coarse-to-fine methods [11], [41] design a two-stage network, where the first network produces an initial coarse prediction, and the second network predicts refined results with the inputs from the first network. The common drawback of these methods is that they focus on information from surrounding pixels rather than reasonable object structures, which result in oversmoothing or blurring. To handle this, [16], [44], [45] proposed a kind of contour-guided model that uses a contour completion network to predict reasonable contour guidance for substantial image inpainting.

B. Face Inpainting

As face editing has become popular in people's daily lives, many researchers have designed specialized methods [11], [16], [43], [44], [46], [47] to solve face inpainting. Early methods [17], [18] reconstructed holes with unchanged face structures by adding semantic parsing loss or perceptual loss. The geometry-aware method [19] proposed a geometry-aware inpainting model that uses a facial geometry estimator to predict a facial geometry from the unmasked region to guide face inpainting and face editing. The recurrent network [48] inserts an RNN model between two CNN models to handle the problem of missing high-frequency details. Furthermore, [4] proposed a multiscale neural patch synthesis model with the concatenation of latent vectors to support high-resolution face completion and attribute control. Although these methods maintained facial structure consistency, they all failed to perceive the identity information from the uncorrupted region. Early ExGAN [1] solves eye-inpainting tasks by using an exemplar image or an encoded code to produce personalized inpainting results, but it produces results of bad global structure and style consistency. Zhao *et al.* [20] proposed an identity-preserving face completion model that takes inference images and pose code to guide identity-preserving inpainting; however, it depends on the pose code and fails to handle large-size images. The recent method proposed by [49] takes the identity knowledge from a pretrained recognizer to supervise the training process, but it cannot fill local holes (e.g., eye inpainting) with the guidance of different identities and cannot handle large-size images as well. As the inpainting task is heavily conditioned by unoccluded regions, these identity-preserving methods are trained to preserve the identity information of the same people. When filling holes, especially partial holes, with the guidance of cross-identity images, using identity information extracted from the entire face always leads to poor results.

C. Face Swapping

Face swapping changes the facial identity of the target face with the identity from the source face while keeping the attributes (e.g., pose, expression, etc.) of the target face unchanged. Early efforts [50], [51] can only exchange faces with similar poses. The 3D-based approaches [52]–[55] transfer identity by matching a 3D morphable face model. As they hardly leverage attributes such as lighting or photo styles, they improve the consistency of pose and expression but lead to artifacts on the swapped faces. Recent GAN-based algorithms [22]–[24], [56], [57] greatly improve the fidelity of swapped images. GAN-based algorithms can be roughly divided into two categories: subject-specific approaches [21], [22], [58]–[60] and subject-agnostic approaches [23]–[26], [56], [61]. The first category takes effect between pairs of subjects and requires a large number of subject-specific images to learn the potential correspondence. In this category, [22], [60] solved the swapping problem with the idea of style transfer, by which it renders the target image in the style of the source image. The popular project DeepFakes [21] trained a specific model for every pair, with thousands of images from two identity folders. Unlike the subject-specific category, there is no need to train a new model for any new face pair for the subject-agnostic approaches. RSGAN [56] first extracts the latent embeddings of the face and hair region separately and then combines the two to synthesize a new face with the help of a foreground mask. IPGAN [25] and FaceShifter [24] implement face swapping by designing a special generator to integrate the identity from the source face and the attributes from the target face. A unified application [61] extracts more attributes from the target and source faces, i.e., the identity, pose/expression and style attributes. The latest method, FaceInpainter [26], uses the idea of facial inpainting to solve the problem of face swapping; it fills in missing areas by fusing identity codes from the source image and the posture codes from the target face. At the same time, we noticed that some image synthesis techniques [62]–[64] also achieved excellent performance in face swapping. Recent studies [27], [65], [66] start to discuss solving the problem by transferring facial attributes at instance level from exemplars. For example, [27] learns the feature embeddings of every face component (e.g., mouth, hair, eye and skin) from the source image, and then generates a new face by the use of a facial mask from the target face.

From a certain point of view, we can treat face swapping as a full-face inpainting problem, where the hole covering the entire face is filled with the identity information from the reference face, and a large proportion of methods [23], [26], [56], [61] mentioned above handle the problem with the same strategy. However, they must depend on the posture or expression attributes from the ground truth; when the pose or expression attribute is missing, e.g., the incomplete image, they fail to generate structurally correct results; they present poor performance in swapping partial faces. Our SwapInpaint combines the advantages of both image inpainting and face swapping to solve the challenging problem of identity-specific face inpainting, especially partial-face inpainting.

III. Method

The identity-specific inpainting task is supposed to fill holes with content of identity modification and structure correctness. In this section, we first introduce the definition of the problem. Then, we introduce the submodules and the detailed architecture. In the following sections, a more detailed experimental setup and comparisons are presented.

A. Problem Formulation

Given an incomplete face image and a reference image, our goal is to complete the incomplete image under the guidance of the reference image. Unlike some early methods that copy pixels directly from the reference, our method learns to fill the missing region with plausible and identity-specific content. The overall framework of our inpainting model is shown in Figure 3. It consists of three modules: identity encoding module I, content inference module C, and generative module G. We define the ground truth image as x_i, the reference image as r_i, the correspondence mask for each incomplete face as m_i and the variable s_i to indicate whether the ground truth of x_i and r_i is the same. Then, we define the training set as $\{X, R, M, S\}$, where $X = \{x_1, x_2, \cdots, x_n\}$, $R = \{r_1, r_2, \cdots, r_n\}$, $M = \{m_1, m_2, \cdots, m_n\}$, and $S = \{s_1, s_2, \cdots, s_n\}$. In addition, we define the incomplete images cropped by masks as $\tilde{X}$. First, we extract the identity code z_{id} through the identity encoding module I by $z_{id} = I(R)$ and infer the content image $\hat{X}$ through the content inference module C by $\hat{X} = C(\tilde{X}, M)$; then, both the identity code z_{id} and content image $\hat{X}$ are input into the generative module to generate the swapped output Y_{sp} by $Y_{sp} = G(\hat{X}, z_{id}, M)$, and the training output Y by a cropping operation. The training process comprise of two stages. First, we train the module C. Then, we put the outputs from modules I and C into module G. For module G, modules I and C are pretrained models, which are included in both the training and predicting phases. The blending operation is only included in the prediction process. As the masks M are used to set the blending boundaries, we define the blending operation as $f_b(\tilde{X}, Y_{sp}, M)$. In addition, a discriminator D is included in the training process to produce realistic results.

B. Content Inference Module

Because that swapping network swaps faces based on full faces and cannot be used in partial-face inpainting directly, we propose a content inference module to predict the missing content. This module can be borrowed from the frontier general inpainting networks, which take incomplete images $\tilde{X}$ and masks M as input and produce the output of inpainted results $\hat{X}$. From the viewpoint of the generative model, we treat the inpainting as a conditional generative problem, which is to learn a face distribution on the condition of the incomplete image and the mask, i.e., $P(\hat{X}|\tilde{X}, M)$. Both autoencoders and VAEs are popular networks for learning the distribution. To improve the fidelity, a GAN discriminator is combined with the networks. Our method supports both autoencoder-based models and VAE-based models in the content inference module as long as they can produce correct content in the facial structure.

Autoencoder-based networks consist of an encoder and a GAN. The encoder encodes the incomplete image $\tilde{x}_i$ into low-dimensional vector z_i, which is then input into the GAN structure to generate a photorealistic result $\hat{x}_i$. The final loss can be the addition of GAN loss and reconstruction loss, which is defined as:

$$L = L_{gan} + \lambda L_{rec}, \tag{1}$$

where the GAN loss is similar to the standard GAN formulation in [15], such as:

$$\begin{aligned} L_{gan}(G, D) = E_{x_i \sim p_{data}(x)} \left[\log D(x_i)\right] \\ + E_{z_i \sim P_z(z)} \left[\log 1 - D(G(z_i))\right], \end{aligned} \tag{2}$$

where the reconstruction loss measures the distance of the generated result to the ground truth image. Here, we choose L1 loss because it produces sharper results compared with L2 loss in inpainting tasks. The reconstruction loss is defined as:

$$L_{rec} = \left\| \hat{x}_i - x_i \right\|_1. \tag{3}$$

Compared with autoencoder-based networks that generate deterministic results, VAE-based networks can produce diverse outputs by sampling from a latent distribution. In addition to the reconstruction loss, a KL loss is used to minimize the gap between the learned latent distribution and the standard normal distribution. Then, the final loss updates as:

$$L = L_{gan} + \lambda_1 L_{rec} + \lambda_2 L_{KL}, \tag{4}$$

$$L_{KL} = -KL\left(P_\phi(z|\tilde{x}_i) || N(0, I)\right), \tag{5}$$

where $P_\phi(.|\tilde{x})$ is the learned latent distribution and $N(0, I)$ is the Gaussian distribution.

C. Generative Module

Inspired by [24], [25], we build a generative module G based on a face-swap network, which disentangles attribute embeddings (e.g., identity, pose, and expression, etc.), and then recombines them to produce a new face with a different identity but the same pose and expression. Since these methods cannot be applied in inpainting tasks directly, we use the intermediate results from the content inference module to supplement the content information to solve the problem of identity-specific inpainting.

Our generative module G requires four inputs, i.e., the identity vectors z_{id}, content inference images $\hat{X}$, masks M, and the same variables S. The identity vectors z_{id} are extracted from a pretrained identity encoding module I, and the content inference images $\hat{X}$ are the intermediate outputs from the content inference module C. In the generative module, we use a subnet A to extract the attribute embeddings z_{att} and a generator to combine z_{id} and z_{att} to produce swapped faces Y_{sp}. The learning objective is to minimize the distance of reference identity $I(R)$ and swapped identity $I\left(Y_{sp}\right)$ and the gap between content attributes $A\left(\hat{X}\right)$ and swapped attribute $A\left(Y_{sp}\right)$. After the addition of GAN loss and reconstruction loss, the optional final loss can be defined as:

$$L = L_{gan} + \lambda_1 L_{rec} + \lambda_2 L_{id} + \lambda_3 L_{att}, \tag{6}$$

where L_{id} is measured by the cosine similarity of two vectors as:

$$L_{id} = 1 - \cos\left(I\left(Y_{sp}\right), I(R)\right), \tag{7}$$

and L_{att} is defined as the $L2$ distance between attribute embeddings as:

$$L_{att} = \frac{1}{2}\left\| A\left(Y_{sp}\right) - A\left(\hat{X}\right) \right\|_2^2. \tag{8}$$

L_{rec} is conditioned by the variables S; when the ground truth image and the reference image are the same, L_{rec} is measured by pixel level $L2$ distance of Y_{sp} and $\hat{X}$. However, the $\hat{X}$ have been redrawed by the module C, we redefine the L_{rec} as:

$$L_{rec} = \begin{cases} \frac{1}{2}\left\| M.Y_{sp} - M.\hat{X} \right\|_2^2 & if\ S = 1; \\ 0 & otherwise. \end{cases} \tag{9}$$

Although the pose attributes of the output image match well with those of the content image, the difference in expression is still evident. In our work, we propose a double attribute loss to minimize the expression mismatch. We crop the masked region from the swapped image and then paste it onto the inference inputs to obtain training outputs Y. According to the task requirements, the attributes of Y are still consistent with the inputs, so we update the attribute loss as:

$$L_{att} = \frac{1}{2}\left(\left\| A\left(Y_{sp}\right) - A\left(\hat{X}\right) \right\|_2^2 + \left\| A(Y) - A\left(\hat{X}\right) \right\|_2^2\right). \tag{10}$$

D. Model Architecture

As depicted in Figure 3, our system disentangles the identity-specific inpainting as content inference and face swapping. For all experiments, we construct a VAE-based content inference network similar to the network proposed in [12] and build a new generative network according to the introduction of [24]. In addition, some necessary modifications are made to support the training for this identity-specific inpainting task.

1) Content Inference Module: Since the content inference module is used to infer reasonable content for the missing areas, it does not require that much diversity. To increase the training speed, we adopt the one-path network, which has a smaller model size. We retain the improvements to the residual block in [12], i.e., replacing the batch normalization in the decoder network with instance normalization and removing the batch normalization in other networks. To preserve structure details, we restrict the training with a multilevel reconstruction loss. The ground truth image is multiscaled to $\left\{x_i^1, x_i^2, \cdots, x_i^n\right\}$, and the outputs $\left\{\hat{x}_i^1, \hat{x}_i^2, \cdots, \hat{x}_i^n\right\}$ are produced from a U-Net decoder, where $\hat{x}_i^k$ indicates the k-th level output. The reconstruction loss is the sum of losses of the output layers. The updated reconstruction loss is:

$$L_{rec} = \sum_{k=1}^{n} \left\| \hat{x}_i^k - x_i^k \right\|_1, \tag{11}$$

where the k is set to 4.

In addition, to make better use of the LeakyReLU functions, we normalize the training images to $[-1, 1]$ and use the tanh

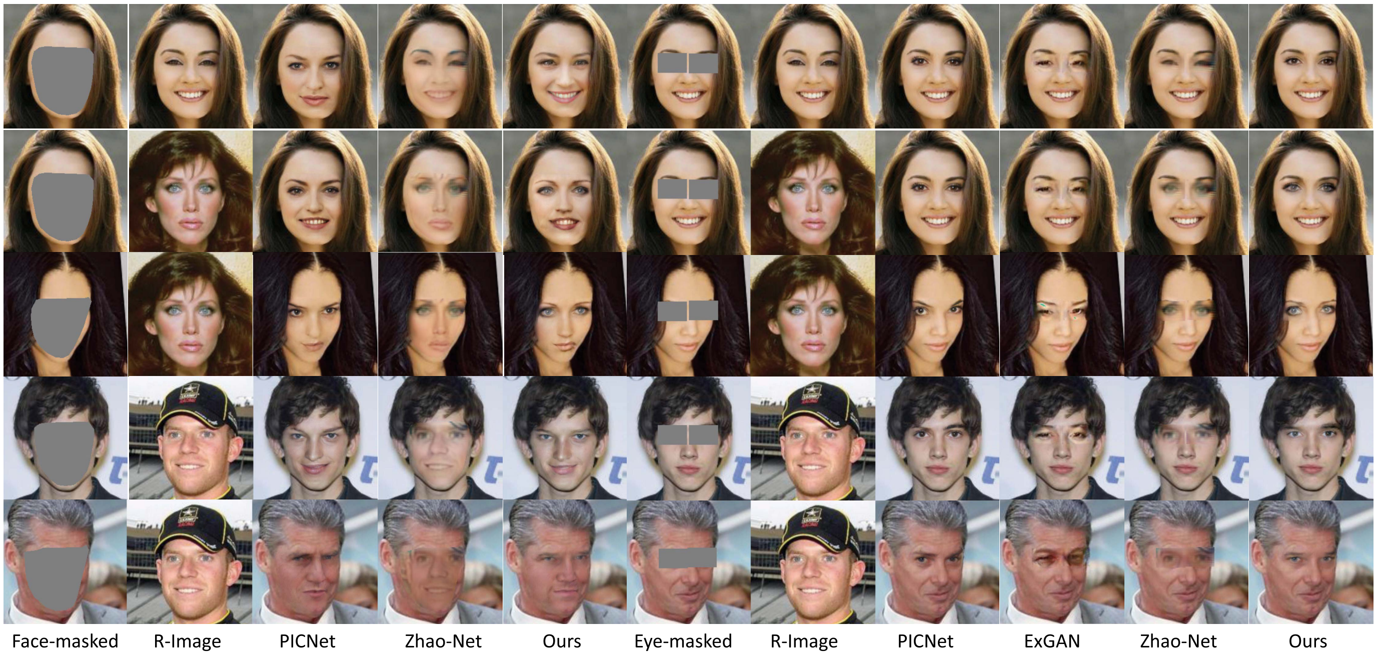

Fig. 4. Comparison of face inpainting tasks with PICNet [12], ExGAN [1] and Zhao-Net [20]. Our model achieves higher identity similarity and structural correctness.

TABLE I
COMPARISON OF MSE, SSIM, PSNR, FID, IDENTITY, POSE, AND EXPRESSION METRICS

Task	Method	MSE↓	SSIM↑	PSNR↑	FID↓	Identity↑	Pose↓	Expression↓
Face	ExGAN	–	–	–	–	–	–	–
	Zhao-Net	221.9588	0.8748	25.3347	23.3179	0.4117	7.6370	71.7862
	Gu-Net	256.6627	0.8691	24.5583	31.8030	0.2893	3.2667	25.6745
	Ours	**243.0506**	**0.8793**	**24.9127**	**13.1548**	**0.4145**	**5.7328**	**52.4547**
Eye	ExGAN	139.2274	0.9521	27.4379	7.9375	0.2596	2.5191	24.0486
	Zhao-Net	111.5892	0.9501	29.2190	19.7196	0.2860	3.7285	33.0287
	Gu-Net	53.3331	0.9440	30.4318	5.6093	0.2756	1.5467	17.9683
	Ours	**80.2452**	**0.9608**	**29.6432**	**5.0090**	**0.2906**	**1.9756**	**22.3370**

function in the last output layer. Moreover, to accommodate partial inpainting, irregular "holes" must be supported. The processing of irregular masks here is referred to [40].

2) Identity Encoder : We use a pretrained state-of-the-art face recognition model [67] as the identity encoder, which is built by ResNet101 with the "num_classes" parameter of 256. It loads a model file trained on VGGFace2 [68], takes an input image of size (112, 112), and produces an identity vector with a length of 256. In addition, all identity loss is computed by the cos distance.

3) Generative Module: Inspired by [24], we propose a subject-agnostic face-swap network to generate swapped faces. We take 8 AAD blocks to integrate the embeddings of identity and attributes in the generator networks as [24] does. To be consistent with the content inference module, we use the LeakyReLU function in the attribute encoder subnet and the tanh function in the output layer of the generator.

4) Blending Operation: We use Poisson blending [69] to remove the style discontinuities between the inpainted contents from Y_{sp} and the background contents from $\hat{X}$ at the end of the prediction process. Masks M are also needed to set the blending boundaries during the blending process.

There are two training strategies. The first trains modules C and G separately in the training process and inputs the inference image from C to the G module to generate the training output. The second module trains the C module first and combines the pretrained I and C to train the improved G module. Both strategies require a blending operation to eliminate the style difference. We experimented with the two approaches and found that the second one presented a better reality, so we fixed our method as the second set.

IV. EXPERIMENT AND RESULTS

In this section, we first compare our method with the state-of-the-art works in the tasks of face inpainting and face swapping. Both full-face inpainting and partial-face inpainting experiments are conducted in each kind of tasks to demonstrate our method's ability to inpaint face images by properly structured and identity-specific content. Then, we analyze the framework to verify the necessity of the inference module and the difference between the two training strategies. Finally, additional examples are presented to demonstrate the generalization ability of our method.

A. Datasets and Experimental Settings

We validate the effectiveness of our proposed method on face inpainting and face swapping tasks, separately. For both tasks, we train our model with CelebA [70], the same as other comparison models. However, we use a larger dataset including CelebA, FFHQ [71], and VGGFace2 [68] to analyze the generalization ability. We set $\lambda_1 = 20$ and $\lambda_2 = 20$ for training the content inference module in formula 4 and set $\lambda_1 = 10$, $\lambda_2 = 5$, and $\lambda_3 = 10$ for the generative module in formula 6.

For all the training and testing images, we use dlib [72] to detect the face regions, followed by aligning and cropping to obtain the 256 × 256 -resolution datasets. Four types of masks are available: center, random, facial and ocular. The center mask is a square region in the middle with a width of three-fifths by face image. The random masks are generated by the method proposed in Pathak *et al.* [40]. Facial and ocular masks are generated according to the face landmarks extracted by dlib. To support random shape hole filling, we use masks consisting of both center and random types to train the generic inpainting method, Zhao-Net and our proposed method. ExGAN is trained with ocular masks as demanded.

We implement our model with PyTorch 1.7.1, train our model on a single GPU, with a batch size of 16 on an NVIDIA V100. All networks are trained from scratch with a fixed learning rate of 10^{-4}. We run training for 2×10^5 iterations on CelebA. After training, the inpainting speed can reach a rate of four per minute.

B. Comparison With Existing Works

1) Qualitative Comparison: We make qualitative comparisons with state-of-the-art methods on the identity-specific face inpainting and face swapping tasks.

a) Face inpainting: Face inpainting deals with the incomplete images. We make comparisons with PICNet [12], ExGAN [1] and Zhao-Net proposed by Zhao *et al.* [20] on cross-image pairs from CelebA. Both full-face inpainting and partial-face inpainting experiments are conducted. From the examples shown in Figure 4, we can see that PICNet produces results that are far from the ground truth but maintain high structural consistency. Zhao-Net depends on pose maps and produces content with inconsistent structures when the pose difference is large. ExGAN performed slightly better than Zhao-Net in eye inpainting. Compared with these methods, our method achieves higher identity similarity and structure correctness.

b) Face swapping: We compare our method with Gu-Net [27] in the face swapping task. For full-face swapping, we take the components of eyes, mouth, and skin from the reference image, deform them to fit the target mask. For eye swapping, we only take the eye components from the reference image. Since the input of swapping tasks is a complete image (cp. incomplete input in inpainting task), it already offers the structure and style information, which help Gu-Net achieve better posture and expression consistency. However, it tends to lose identity information when the facial features are too deformed, as shown in Figure 5 and Table I.

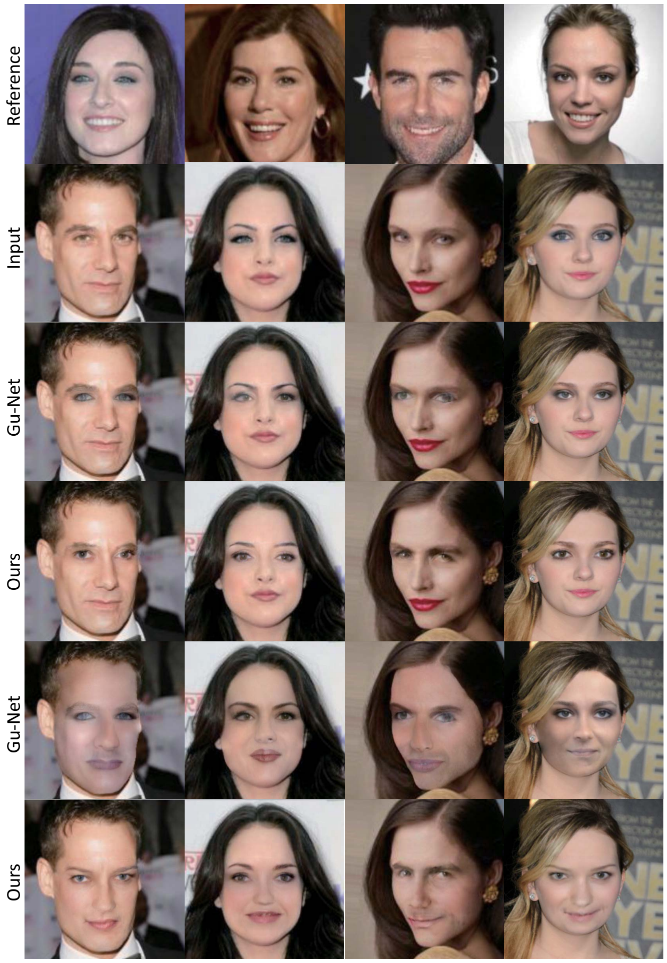

Fig. 5. Comparison with Gu-Net [27] on eye and full-face swapping. The third and fourth lines are eye swap, while the fifth and sixth lines are face swap. Our model achieves higher fidelity and identity similarity.

2) Quantitative Comparison: It is difficult to quantize inpainting tasks, but we can measure the structural consistency and identity similarity through some metrics. We measure pixel-level similarity through mean squared error (MSE), structural similarity index (SSIM), and the peak signal-to-noise ratio (PSNR), weigh the quality of generated images with the Fréchet inception distance (FID) [37], evaluate identity similarity with the identity retrieval score, and compare the structural consistency using pose and expression, as shown in Table I.

For image similarity, the MSE, SSIM and PSNR are computed between the ground truth and inpainted result. The FID score is computed between the groups of ground truth images and inpainted images. For identity-filling capability, we use CosFace [73] trained on CAISA-WebFace [74] to extract identity vectors and compare the identity similarity between the reference image and inpainted result by cosine distance. For structural consistency, head pose and expression landmarks are retrieved from a pretrained HopeNet [75]

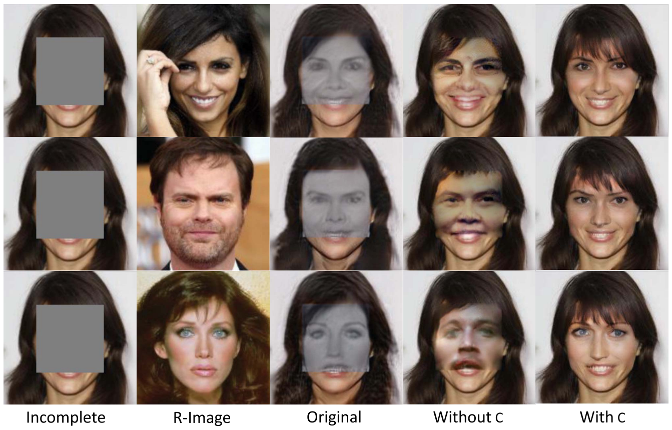

Fig. 6. Comparison with two baseline models. The original AEI-Net is trained with complete image pairs but takes incomplete images when inpainting. The simply improved AEI-Net without a content inference module (C) is trained with incomplete image and reference image pairs.

TABLE II

COMPARISON ON METRICS OF SSIM, FID, IDENTITY, POSE AND EXPRESION

Method	SSIM↑	FID↓	Identity↑	Pose↓	Exp.↓
Original	0.8277	109.8328	0.2549	9.6395	60.8501
Without C	0.8379	20.6367	0.2900	8.3263	65.4460
With C	**0.8736**	**13.7581**	**0.3044**	**5.4274**	**50.2071**

and FAN detector [76], respectively, and compared with the absolute distance between the ground truth and the inpainted face.

The experiments are all constructed on 1k test images of cross pairs from CelebA. We perform both full-face inpainting and eye-inpainting experiments on all models except the ExGAN. Since the ExGAN is only available for eye inpainting, only eye-inpainting experiments are conducted.

As the table shows, Our method achieves a higher degree of identity similarity and image fidelity on both full and partial identity-specific face inpainting tasks.

C. Analysis of the Framework

1) Content Inference Module: To analyze the necessity of the content inference module, we compare our model with two baseline models: 1) original AEI-Net trained by complete image pairs and 2) simple improved AEI-Net with incomplete image and reference image pairs. We compare the two baseline models and our method in Figure 6. When a masked face is input into the original AEI-Net, the outputs miss style information. The simple improved model can produce complete results but with structure inconsistency. Our model uses a content inference module to supplement attribute information, which enables high qualitative inpainting. Table II shows the quantitative comparison of the metrics of SSIM, FID, identity, pose and expression. We conduct experiments on cross pairs with center mask. As shown in the table, our method outperforms the other two baselines by large margins.

2) Contextual Loss: FaceInpainter [26] implemented a styled face inpainting network by using the contextual loss [77], which is used to reduce texture distortion and eliminate the need for blending processing. We define the contextual loss as L_{CX} and compute it upon the facial content extracted by a pretrained face parsing model [78], using the ReLU{3_2, 4_2} layers of the pretrained VGG19 network as the FaceInpainter. We utilize the contextual loss in our training

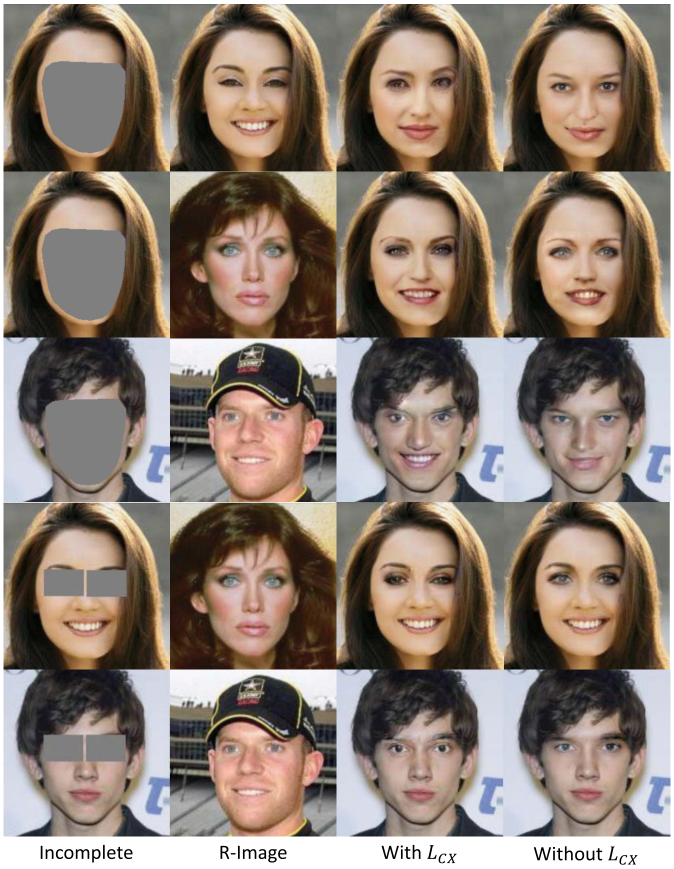

Fig. 7. Comparison between cases with L_{CX} and without L_{CX}.

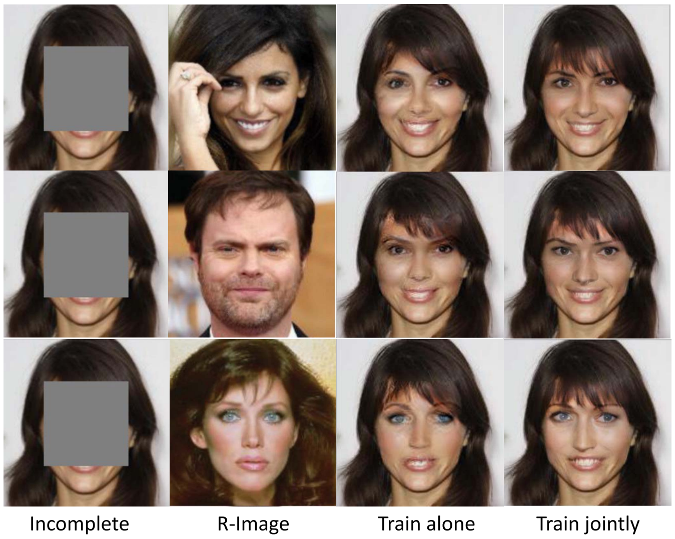

Fig. 8. Comparison with two training strategies: training alone and training jointly. The second strategy performs better.

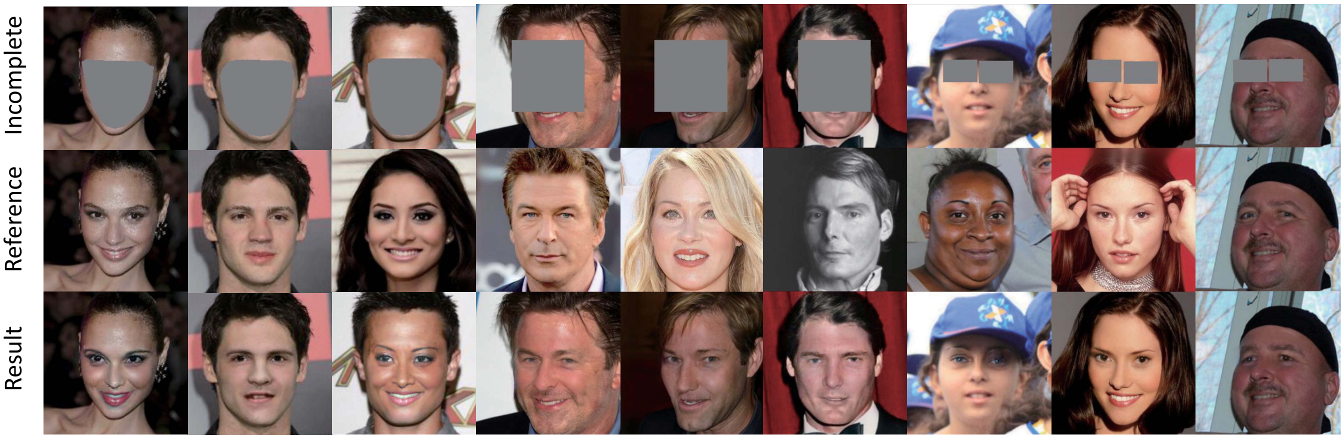

Fig. 9. Our inpainting results on wild reference face images.

process and find that the style consistency improved, but some identity features (e.g., eye color) vanished, as shown in Figure 7.

3) Training Strategy: We compare two training strategies, i.e., train alone and train jointly. We experimented with the two strategies in Figure 8 and found that the second strategy achieves better attribute consistency.

4) Generalization Ability: Finally, we demonstrate the generalization ability of our method by testing on wild images, which is shown in Figure 9.

V. Conclusion

We proposed SwapInpaint, a new identity-specific face inpainting model for image inpainting tasks. Our model combines a subject-agnostic face-swap network with a face content inference network and affects both full and partial-face inpainting. Through comparison with existing inpainting methods and basic AEI-Net, we demonstrate that the results of our method outperform the state-of-the-art results and that the content inference module plays an important role in identity-specific inpainting tasks.

Although our proposed methods produce high-quality results, there is still plenty of room for improvement. For example, the choice of the Poisson blending algorithm in the blending module greatly slows down the predicting speed. In addition, we found that there are still problems about how to judge the optimal tradeoff point to minimize identity loss while keeping the attributes consistent with the incomplete input.

VI. Future Work

As the aforementioned limitations, we plan to investigate the solution of increasing the prediction speed, make more analysis on the overfitting problem. Future work also includes extending our work to video inpainting applications.

References

[1] B. Dolhansky and C. C. Ferrer, “Eye in-painting with exemplar generative adversarial networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7902–7911.

[2] M. Brand and P. Pletscher, “A conditional random field for automatic photo editing,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.

[3] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, “Digital face beautification,” in *Proc. ACM SIGGRAPH Sketches*, Jul. 2006, p. 169.

[4] Z. Chen, S. Nie, T. Wu, and C. G. Healey, “High resolution face completion with multiple controllable attributes via fully end-to-end progressive generative adversarial networks,” 2018, *arXiv:1801.07632*.

[5] Q. Duan, L. Zhang, and X. Gao, “Simultaneous face completion and frontalization via mask guided two-stage GAN,” *IEEE Trans. Circuits Syst. Video Technol.*, early access, Sep. 10, 2021, doi: 10.1109/TCSVT.2021.3111648.

[6] Z. Wan *et al.*, “Bringing old photos back to life,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2747–2757.

[7] X. Tu *et al.*, “Joint face image restoration and frontalization for recognition,” *IEEE Trans. Circuits Syst. Video Technol.*, early access, May 10, 2021, doi: 10.1109/TCSVT.2021.3078517.

[8] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, “Face recognition by humans: Nineteen results all computer vision researchers should know about,” *Proc. IEEE*, vol. 94, no. 11, pp. 1948–1962, Nov. 2006.

[9] S. Iizuka, E. Simo-Serra, and H. Ishikawa, “Globally and locally consistent image completion,” *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–14, 2017.

[10] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *Proc. 36th Int. Conf. Mach. Learn.*, Jun. 2019, pp. 7354–7363.

[11] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, “Free-form image inpainting with gated convolution,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.

[12] C. Zheng, T.-J. Cham, and J. Cai, “Pluralistic image completion,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1438–1447.

[13] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” in *Proc. Int. Conf. Artif. Neural Netw.* Springer, 2011, pp. 52–59.

[14] D. P. Kingma and M. Welling, “Auto-encoding variational Bayes,” 2013, *arXiv:1312.6114*.

[15] I. Goodfellow *et al.*, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014.

[16] W. Xiong *et al.*, “Foreground-aware image inpainting,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5840–5848.

[17] Y. Li, S. Liu, J. Yang, and M.-H. Yang, “Generative face completion,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3911–3919.

[18] R. Ma and H. Hu, “Perceptual face completion using a local-global generative adversarial network,” in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1670–1675.

[19] L. Song, J. Cao, L. Song, Y. Hu, and R. He, “Geometry-aware face completion and editing,” in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 2506–2513.

[20] Y. Zhao *et al.*, “Identity preserving face completion for large ocular region occlusion,” 2018, *arXiv:1807.08772*.

[21] L. Floridi, “Artificial intelligence, deepfakes and a future of ectypes,” *Philosophy Technol.*, vol. 31, no. 3, pp. 317–321, 2018.

[22] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast face-swap using convolutional neural networks,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3677–3685.

[23] Y. Nirkin, Y. Keller, and T. Hassner, “FSGAN: Subject agnostic face swapping and reenactment,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7184–7193.

[24] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, “FaceShifter: Towards high fidelity and occlusion aware face swapping,” 2019, *arXiv:1912.13457*.

[25] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, “Towards open-set identity preserving face synthesis,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6713–6722.

[26] J. Li, Z. Li, J. Cao, X. Song, and R. He, “FaceInpainter: High fidelity face adaptation to heterogeneous domains,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5089–5098.

[27] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, “Mask-guided portrait editing with conditional GANs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3436–3445.

[28] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, “Image inpainting,” in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, Jul. 2000, pp. 417–424.

[29] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, “PatchMatch: A randomized correspondence algorithm for structural image editing,” *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[30] M. Bertalmio, L. Vese, G. Sapiro, and S. Osher, “Simultaneous structure and texture image inpainting,” *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 882–889, Aug. 2003.

[31] A. Criminisi, P. Perez, and K. Toyama, “Object removal by exemplar-based inpainting,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2003, pp. 1–8.

[32] A. Criminisi, P. Pérez, and K. Toyama, “Region filling and object removal by exemplar-based image inpainting,” *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1200–1212, Sep. 2004.

[33] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, “Image completion using planar structure guidance,” *ACM Trans. Graph.*, vol. 33, no. 4, pp. 1–10, Jul. 2014.

[34] H.-Q. Wang, Q. Chen, C.-H. Hsieh, and P. Yu, “Fast exemplar-based image inpainting approach,” in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 5, Jul. 2012, pp. 1743–1747.

[35] N. Gopinath, K. Arjun, J. A. Shankar, and J. J. Nair, “Complex difffusion based image inpainting,” in *Proc. 1st Int. Conf. Next Gener. Comput. Technol. (NGCT)*, Sep. 2015, pp. 976–980.

[36] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, “Semantic image inpainting with deep generative models,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5485–5493.

[37] J. Zhang *et al.*, “GAIN: Gradient augmented inpainting network for irregular holes,” in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1870–1878.

[38] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, “Context encoders: Feature learning by inpainting,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2536–2544.

[39] A. Lahiri, S. Bairagya, S. Bera, S. Haldar, and P. K. Biswas, “Lightweight modules for efficient deep learning based image restoration,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1395–1410, Apr. 2021.

[40] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, “Image inpainting for irregular holes using partial convolutions,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 85–100.

[41] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, “Generative image inpainting with contextual attention,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[42] L. Zhao *et al.*, “UCTGAN: Diverse image inpainting based on unsupervised cross-space translation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5741–5750.

[43] A. Lahiri, A. K. Jain, S. Agrawal, P. Mitra, and P. K. Biswas, “Prior guided GAN based semantic inpainting,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13696–13705.

[44] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, “EdgeConnect: Generative image inpainting with adversarial edge learning,” 2019, *arXiv:1901.00212*.

[45] S. Xu, D. Liu, and Z. Xiong, “E2I: Generative inpainting from edge to image,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1308–1322, Apr. 2021.

[46] H. Zhang and T. Li, “Semantic face image inpainting based on generative adversarial network,” in *Proc. 35th Youth Acad. Annu. Conf. Chin. Assoc. Autom. (YAC)*, Oct. 2020, pp. 530–535.

[47] B. Jiang, H. Liu, C. Yang, S. Huang, and Y. Xiao, “Face inpainting with dilated skip architecture and multi-scale adversarial networks,” in *Proc. 9th Int. Symp. Parallel Archit., Algorithms Program. (PAAP)*, Dec. 2018, pp. 211–218.

[48] Q. Wang, H. Fan, G. Sun, W. Ren, and Y. Tang, “Recurrent generative adversarial network for face completion,” *IEEE Trans. Multimedia*, vol. 23, pp. 429–442, 2020.

[49] S. Ge, C. Li, S. Zhao, and D. Zeng, “Occluded face recognition in the wild by identity-diversity inpainting,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3387–3397, Oct. 2020.

[50] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, “Face swapping: Automatically replacing faces in photographs,” in *Proc. ACM SIGGRAPH Papers*, Aug. 2008, pp. 1–8.

[51] H.-X. Wang, C. Pan, H. Gong, and H.-Y. Wu, “Facial image composition based on active appearance model,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 893–896.

[52] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, “Exchanging faces in images,” in *Computer Graphics Forum*, vol. 23. Hoboken, NJ, USA: Wiley, 2004, pp. 669–676.

[53] Y.-T. Cheng *et al.*, “3D-model-based face replacement in video,” in *Proc. SIGGRAPH, Posters*, Aug. 2009, p. 1.

[54] Y. Lin, S. Wang, Q. Lin, and F. Tang, “Face swapping under large pose variations: A 3D model based approach,” in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 333–338.

[55] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, “On face segmentation, face swapping, and face perception,” in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 98–105.

[56] R. Natsume, T. Yatagawa, and S. Morishima, “RSGAN: Face swapping and editing using face and hair representation in latent spaces,” 2018, *arXiv:1804.03447*.

[57] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, “CVAE-GAN: Fine-grained image generation through asymmetric training,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2745–2754.

[58] W. Wu, Y. Zhang, C. Li, C. Qian, and C. C. Loy, “ReenactGAN: Learning to reenact faces via boundary transfer,” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 603–619.

[59] H. Kim *et al.*, “Deep video portraits,” *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–14, Aug. 2018.

[60] X. Jin, Y. Qi, and S. Wu, “CycleGAN face-off,” 2017, *arXiv:1712.03451*.

[61] L. M. Ngo, C. A. de Wiel, S. Karaoglu, and T. Gevers, “Unified application of style transfer for face swapping and reenactment,” in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2020, pp. 1–17.

[62] J.-Y. Zhu *et al.*, “Multimodal image-to-image translation by enforcing bi-cycle consistency,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 465–476.

[63] D. Y. Park and K. H. Lee, “Arbitrary style transfer with style-attentional networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5880–5888.

[64] H. Zheng, H. Liao, L. Chen, W. Xiong, T. Chen, and J. Luo, “Example-guided image synthesis using masked spatial-channel attention and self-supervision,” in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 422–439.

[65] Q. Deng, J. Cao, Y. Liu, Z. Chai, Q. Li, and Z. Sun, “Reference-guided face component editing,” 2020, *arXiv:2006.02051*.

[66] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “MaskGAN: Towards diverse and interactive facial image manipulation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5549–5558.

[67] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “ArcFace: Additive angular margin loss for deep face recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[68] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[69] P. Pérez, M. Gangnet, and A. Blake, “Poisson image editing,” in *Proc. ACM SIGGRAPH Papers*, Jul. 2003, pp. 313–318.

[70] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[71] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[72] D. E. King, “Dlib-ml: A machine learning toolkit,” *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, Jan. 2009.

[73] H. Wang *et al.*, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

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

[75] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2074–2083.

[76] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (And a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[77] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 768–783.

[78] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 325–341.

Honglei Li received the B.S. and M.S. degrees in software engineering from Chongqing University, Chongqing, China, in 2010 and 2012, respectively. She is currently pursuing the Ph.D. degree in artificial intelligence with the Macau University of Science and Technology, Macau.

From 2014 to 2017, she was a Research Assistant with the Laboratory of Software, College of Computer Science, Chongqing University. After that, she worked as a Lecturer with the Chongqing College of Electronic Engineering, Chongqing, before 2020. Her research interests include computer vision, artificial intelligence, machine learning, and software engineering.

Wenmin Wang (Member, IEEE) received the Ph.D. degree in computer architecture from the Harbin Institute of Technology, China, in 1989. After that, he worked as an Assistant Professor and an Associate Professor with the Harbin University of Science and Technology and the Harbin Institute of Technology. Since 1992, he has gaining approximately 18 years of overseas industrial experience in Japan and America, where he served as a Staff Engineer, a Chief Engineer, and the Software Division General Manager. He returned to the academia of China by the end of 2009 as a Professor with the School of Electronic and Computer Engineering, Peking University, China. In 2019, he joined the International Institute of Next Generation Internet, Macau University of Science and Technology as a Professor. His current research interests include computer vision, multimedia retrieval, artificial intelligence, and machine learning.

Cheng Yu received the B.S. degree in computer science and technology from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2011, and the M.S. degree in technology of computer application from Northeastern University, Shenyang, China, in 2017. He is currently pursuing the Ph.D. degree in artificial intelligence with the Macau University of Science and Technology, Macau, China. His blog at: disanda.github.io. His research interests include computer vision, deep learning, generative models, and related applications.

Shixiong Zhang received the M.S. degree in computer and information systems from the Macau University of Science and Technology, Taipa, Macau, China, in 2016, where he is currently pursuing the Ph.D. degree in artificial intelligence. From October 2016 to October 2017, he was a Research Assistant with the Chu Hai College of Higher Education, Hong Kong, China. His current research interests include event-based vision, object detection, and tracking.