Generative Facial Prior and Semantic Guidance for Iterative Face Inpainting

XIN-YU ZHANG1,2,3, KAI XIE1,2,3, MEI-RAN LI1,2,3, CHANG WEN3,4, AND JIAN-BIAO HE5

1School of Electronic and Information, Yangtze University, Jingzhou 434023, China
2National Demonstration Center for Experimental Electrical and Electronic Education, Yangtze University, Jingzhou 434023, China
3Western Institute, Yangtze University, Karamay 834000, China
4School of Computer Science, Yangtze University, Jingzhou 434023, China
5School of Computer Science and Engineering, Central South University, Changsha 410083, China

Corresponding author: Kai Xie (pami2009@163.com)

This work was supported in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2020D01A131, in part by the Fund of Hubei Ministry of Education under Grant B2019039, in part by the Graduate Teaching and Research Fund of Yangtze University under Grant YJY201909, in part by the Teaching and Research Fund of Yangtze University under Grant JY2019011, in part by the Undergraduate Training Programs for Innovation and Entrepreneurship of Yangtze University under Grant Yz2020057, and in part by the National College Student Innovation and Entrepreneurship Training Program under Grant 202110489003.

ABSTRACT Image inpainting techniques have been greatly improved by relying on structure and texture priors. However, damaged original images or rough predictions cannot provide sufficient texture information and accurate structural priors, leading to a drop in image quality. Moreover, from the perspective of human visual perception, it is important to pay attention to facial symmetry and facial attribute consistency. In this paper, we present a face inpainting system with iteration structure, guided by generative facial priors contained in pretrained GANs and predicted semantic information. Specifically, generative facial priors generated by the GAN inversion techniques introduce sufficient textures and features to assist inpainting; semantic maps are able to provide facial structural information and semantic categories of different pixels for face reconstruction. In particular, we iteratively refine images multiple times, updating semantic maps at each iteration. The Weighted Prior-Guidance Modulation layer (WPGM) is devised for incorporating priors into networks through spatial modulation. We also propose facial feature self-symmetry loss to constrain the symmetry of faces in feature space. Experiments on CelebA-HQ and LaPa datasets demonstrate the superior performance of our model for facial detail and attribute consistency. Meanwhile, under the background of COVID-19, it is worth trying recognition via inpainting to deal with recognition challenges brought by mask occlusion. Relevant experiments show that our inpainting model does help to recognition tasks to a certain degree, with higher accuracy.

INDEX TERMS Face inpainting, semantic prior, generative facial prior, iterative structure, face symmetry, recognition via inpainting.

I. INTRODUCTION

Face inpainting, as a branch of image inpainting, is to repair the damaged face images with reasonable visual information. Face images have some specific properties, such as strong facial topological structure and attribute consistency (e.g., expression, pose, gender, age, ethnicity and make-up). Existing learning-based methods [1]–[5] have succeeded in getting a completed image with reasonable content. However, as for face images, it is difficult for these methods to pay attention to face symmetry, and they suffer conflict of attributes as illustrated in Fig. 1. Obviously, the result generated by [1] has mismatches between the left and right eyes, while our method can produce results with consistent style.

Recent structure-based methods [6]–[8] are built on the two-stage architecture and they use predicted contour or structure of the damaged input to complete the missing areas. These methods surely have good performance, and it shows that structure priors are vital to produce realistic images with accurate face shape. However, these models depend almost
entirely on predicted structure and errors of predicted structures at the first stage will directly cause inpainting failure undoubtedly. Although Xiong et al. [9] use detection and completion module together to predict the contour by separating foreground and background. It cannot outline the edge of face components, so it still does not work for face inpainting. Li et al. [10] predict the edge and contour when inpainting. Specifically, the repaired features are utilized to assist structure prediction while simultaneously using structure to assist inpainting. However, methods above only focus on contour prediction and ignore the semantic content, leading to generating blurry boundary.

To solve the problem, and also since structural priors play an important role in inpainting, we use the semantic map, widely applied in Image Generation [11], Image Restoration [12], Super-Resolution [13], to assist face inpainting. Moreover, we iteratively update semantic maps to avoid the risk of one prediction. Semantic maps contain not only facial contour and shape, but also semantic categories which can guide the contribution of different semantic regions to missing parts. Thus, we leverage semantic prior to learn different distribution of face components to reconstruct face images, which certainly constrain face attribute consistency in expression and pose and avoid face distortion. However, the semantic maps are estimated by the coarse prediction of the input so that errors and inaccurate semantic labels are inevitable, which means that only relying on semantic prior may cause wrong prediction or contours. In addition, semantic prior still contain limited texture information to be used for filling holes.

There are some works combine both structure and texture information to assist inpainting and they do achieve great results. Ren et al. [14] use the structure generator and texture generator with appearance flow to fill holes. Liu et al. [15] extract structure and texture from input images to generate natural results. Ding et al. [16] use the nonlocal texture similarity to generate natural images. However, they cannot infer the structure and texture of the missing area with known information when the missing area is too large or not relevant to known information.

Thus, generative facial prior is used to make up for these deficiencies. Wang et al. [17] find that generative facial priors have abundant and diverse facial textures (also including other various priors such as color and style), so they leverage the priors for face restoration task which is to generate a clear face from a degraded face image. As the generative face produced by pretrained GANs has sufficient priors, their model is able to recover face detail and really get realistic and faithful results. Inspired by their work, we decide to use the rich information of generative facial prior and predicted semantic maps to jointly help face completion. Generative faithful faces contain sufficient textures and have consistent style such as age, gender, ethnicity, make-up. These properties can support the model to further improve the perceptual facial details and give more guidance on the style property consistency. Nevertheless, feeding priors directly into networks as in [7] easily lead to insufficient use because this information will disappear in the deeper layer with the normalization.

To address these issues, we propose the Generative Facial Prior and Semantic Guidance for Iterative Face Inpainting (GSGIFI) network, which is generally coarse-to-fine architecture. We consider inpainting to be a constrained image generation problem, so our idea is to utilize prior knowledge as a guide for conditioned image generation. In order to suit our task, on the basis of SPAED [11], we devise WPGM layer to use spatially adaptive normalization to incorporate two kinds of prior information to guide fine inpainting by modulating feature maps. In a word, we leverage abundant features from generative priors to assist inpainting and also use semantic priors to encourage the model to fill missing areas by itself. To address the issue of generating unsymmetrical face, we design the facial feature self-symmetry loss to constrain the left and right face to be symmetrical in feature space. The experiments show that our method outperforms other state-of-the-arts and can produce more realistic and detailed images with consistent attributes.

In conclusion, our contributions can be summarized as follow:

1. We incorporate generative facial priors and semantic priors as guidance to refine inpainting in a novel way. The former priors contain sufficient textures and consistent style, and semantic priors contain structure and semantic category information, allowing our model to achieve face detail inpainting and attributes consistency.

2. We devise WPGM layer to better add priors into the model to guide inpainting. Considering face symmetry, we design the facial feature self-symmetry loss function to further constrain facial features of left and right face of inpainting results to be coordinated in shape, color, texture, etc.

3. We propose GSGIFI network with designs of evaluation and iteration architecture for inpainting. Experiments show that our method achieves better performance than state-of-the-arts and that it is possible to implement “recognition via inpainting”.

**Motivation:** Firstly, face attributes have strong style consistency in expression, gender, color, make-up etc., while existing methods do not perform very well in these aspects as shown in Fig 1. Although Yang et al. [6] have tried to use distant spatial context and connect temporal feature maps to fulfill this, the style of their results tends to be monolithic. Additionally, it is easy to know that features of the left and right faces should be symmetrical, such as wrinkles around
the corners of the right and left eyes. Secondly, as mentioned earlier, recent inpainting methods mainly rely on structure and texture priors, but their two-stage architecture is prone to inpainting failure. Besides, it is difficult to get accurate predicted edge and rich textures from input images or coarse predictions. Thirdly, during COVID-19, people wear masks which occlude a lot of facial features, making it difficult for traditional recognition technology to up to satisfactory accuracy, like [18]. Therefore, we attempt to explore the possibility of recognition via inpainting, using the inpainting techniques to assist recognition, which can also verify what extent our inpainting model can restore lost facial information.

To clarify the structure of this article, the arrangement is shown as follows. Section II presents some background and works related to our paper. Section III mainly describes the overall architecture and detailed theory of the proposed algorithm. Section IV includes the implementation details, experiment results and comparison, ablation study and limitation. Section V concludes the proposed method.

II. RELATED WORK

Image inpainting has always been a hot topic in the field of computer vision, due to its wide application in image repairing [1], [19], editing [20]–[22], object removal [23], etc. Some deep learning methods [1], [20], [24], [25] have reached great achievement in completing holes with correct content. However, these methods by inferring the content of damaged images still encounter difficulties in reasonable structure and realistic textures. Differently, [26], [27] can repair images with mask awareness.

Progressive image inpainting is becoming mainstream, unlike previous methods of directly repairing the target image to obtain results but taking well-inpainting areas as known information to progressively complete holes. [28], [29] cascade multiple generators to progressively fill holes. Li et al. [25] use recurrent architecture to reason features under a small number of parameters. On this basis, Zeng et al. [30] add iterative feedback and further produce high-resolution images, using updated confidence maps as masks. All of these works combine various information to progressively complete the lost content and achieve great performance. Therefore, in order to avoid problems of two-stage architecture mentioned in the Introduction, we adopt iterative architecture to progressively repair face images.

A. PRIOR-GUIDED INPAINTING

Recently, many works exploited face-specific priors, such as facial landmarks [6], visual structure [8], [10], and attention maps to assist inpainting. Facial landmarks and contours are able to guide completion, but their results are not very well done in terms of detail and texture, and it is difficult to ensure accurate landmarks or structure. Compared with these priors, the semantic map includes not only overall facial contour, but also semantic categories of each pixel, shape and location of face components, which means it can offer more information for inpainting. Li et al. [31] use semantic constraint to further strengthen the harmony between the generated content and existing pixels. [7], [32] input predicted semantic maps into networks to help inpainting. Liao et al. [33] use semantic maps and texture to repair the hole layer by layer. These semantic-based methods demonstrated that semantic priors are effective for inpainting, but how to get accurate semantic maps is still challenging. Instead of predicting once, we instead predict multiple times in iterative inpainting to refine the semantic map. With the help of generative facial prior, it can help predict the missing content to obtain more detailed textures and accurate semantic maps.

B. INPAINTING BASED ON GAN

Since Generative Adversarial Network (GAN) was put forward, it has been widely applied in image generation [34], deep translation [35], etc. There are many image inpainting methods based on GAN network such as [19], [31], [36]. In order to get more coherent and realistic results, [14], [37] use local and global ideas for inpainting. Considering that this idea has poor performance under large missing areas, Xu et al. [39] utilize the edgeness map to fill the holes based on GAN training. Previously, Zhang et al. [40] typically inverted the damaged images into latent codes to directly generate face images, but results tend to blur and distort. In this paper, we take a new approach. Pretrained GANs implicitly contain generative priors, such as StyleGAN2 [34] which is able to produce faithful faces according to latent codes. Wang et al. [17] use the priors to achieve great visual effect on low-quality images. Therefore, we argue it can be a good way to use generative facial priors to produce images with detailed textures.

III. METHOD

A. OVERALL ARCHITECTURE

We will describe the overall architecture of the network, and first we give some explanations of notation. $I_{CN}$ denotes the coarse prediction generated by the coarse network; $S_{CN}$ denotes semantic segmentation maps of $I_{CN}$. Given the decoder of four layers, we denote feature maps from shallow to deep in the decoder of the refinement network as $F^{4}_{RDEC}, F^{3}_{RDEC}, F^{2}_{RDEC}, F^{1}_{RDEC}$. Given a pretrained StyleGAN2 network, $F^{3}_{SGAN}, F^{2}_{SGAN}, F^{1}_{SGAN}$ are feature maps in the generation process separately, corresponding to the scale of $F^{4}_{RDEC}, F^{3}_{RDEC}, F^{2}_{RDEC}$. $I_{RN}, S_{RN}$ are the fine prediction and the semantic map of the refinement network respectively.

Fig. 2 shows our iterative model, generally a coarse-to-fine network, which inputs the result of the coarse prediction into the encoder of refinement inpainting network to be encoded. Then, feature maps will be modulated through a single-layer structure with $S_{CN}$ and $F_{SGAN}$ as priors. We use a pretrained Partial convolution network [1] to make prediction of holes and the last feature map $F^{4}_{RDEC}$ is passed through two branches to get the coarse prediction $I_{CN}$ and $S_{CN}$, respectively. The refinement network refines the coarse prediction.
FIGURE 2. Overview of the proposed architecture. (a) We first get a coarse prediction by pretrained Partial UNet [14] and corresponding semantic map first. (b) The coarse prediction is mapped into the latent codes to be input into pretrained StyleGAN2 to generate intermediate feature maps. The model uses the WPGM layer to combine generative facial priors and semantic priors to modulate feature maps for refinement and then update segmentation maps. (c) We update masks by evaluating the final prediction with new segmentation maps. In the next iteration, the fine prediction of the last iteration and the new mask will be input into refinement network to re-inpaint. (b) and (c) are performed continuously until the set number of iterations is reached. q denotes the preset number of iterations. The symmetry of generated face is constrained by the proposed self-symmetry loss and there are also other loss functions $L_1$, $L_{adv}$, $L_{cross}$, $L_{id}$.

by WPGM layer to generate refinement prediction and its segmentation maps. On the one hand, $I_{CN}$ is mapped into the latent space through gated convolution [20] and several linear layers, and then it is input into the pretrained StyleGAN2 [34] to deploy generative facial priors. On the other hand, the encoded $I_{CN}$ is sent into the decoder where adaptively modulates features using semantic maps and generative priors. At last, pixels with wrong semantic information will be labelled out by evaluation and then go next round of inpainting to get better inpainting results. In order to improve the coordination of inpainting face images, we use the facial feature self-symmetry loss function to constrain generation to make the final prediction symmetric for more natural results. We also utilize adversarial loss, identity preserving loss, and cross entropy loss.

B. PRIOR CONDITION
Prior is at the image level, not the pixel level. Many works exploit priors, such as [8] using structural priors as constraint. However, we take an unusual approach, using semantic priors...
and generative facial priors as guidance to produce detailed textures and restore face perception as much as possible.

1) GENERATIVE FACIAL PRIOR
Owing to the convolution weight of pretrained GANs capturing the distribution of the face, GANs learned rich facial knowledge (such as different eye features) on external dataset. The facial knowledge is what generative facial priors contain. We use pretrained GANs to match the output closest to the input and leverage information of the output to guide inpainting. In order to deploy generative facial priors, we first map the damaged images into the latent codes, and then use pretrained StyleGAN2 to generate output close to the input. We argue that the intermediate feature maps contain enough information so that the final faithful face does not need to be generated, which also reduces part of the computation. Specifically, the input images are convolved to get the encoded vector and then pass-through Multilayer Perceptron (MLP) with several linear layers. Therefore, input images are mapped into the latent space to obtain the latent code $z$ that contains facial semantic properties:

$$z = MLP(GatedUNet(I_{in}))$$  \hspace{1cm} (1)

where $I_{in}$ represents the input to GatedUNet. In the first iteration, the input is $I_{CN}$ and the input mask. In the later iteration, $I_{in}$ represent $I_{RN}$ and corresponding mask. GatedUNet is the variant of UNet [41] architecture, replacing the vanilla convolution with gated convolution. UNet is simple but effective CNN network. Contrast with vanilla convolution treating all pixels as valid pixels, gated convolution can use soft gating to update masks for every channel by dynamic learning, which is more suitable for irregular inpainting. Therefore, GatedUNet is used to encode the input images to extract valid features and get the representation of input images.

Then, the latent code $z$ is input into the pretrained StyleGAN2 to generate the closest face output, scales of which are corresponded with $F_{RDEC}$.

$$F_{\theta}^{l}_{SGAN} = StyleGAN(z, \theta_\theta)$$  \hspace{1cm} (2)

where $\theta_\theta$ is the parameters of the pretrained StyleGAN2 and $l$ represents the scales of $F_{SGAN}$. Generative intermediate feature maps $F_{SGAN}$ can offer more diverse textures which can be used to further modulate $F_{RDEC}$. With the introduction of abundant textures and features from intermediate feature maps, the model is easier to generate detailed and reasonable consequence.

2) SEMANTIC PRIOR
There is category information in the semantic maps helping to distinguish which semantic category a pixel belongs to, thus avoiding aircraft and blurring. Besides, semantic maps also contain face shape, beneficial to generating clear boundary. Information used above is collectively called semantic prior contained in semantic maps. Liao et al. [32] apply multi-scale semantic maps to image inpainting, using the SPADE module to gradually infer image content. Similar to their work, we also use multi-scale semantic maps, but we continuously refine the semantic maps through iteration. We first use segmentation maps of coarse prediction to guide the first refinement inpainting, and then we use the semantic maps of first fine prediction to guide next refinement. Therefore, the semantic maps refinement and inpainting process are simultaneously carried out and the two interact with each other. In the first iteration, the segmentation maps $S_{CN}$ for guiding inpainting can be easily obtained, similar to [32]:

$$S_{CN} = S(F^0)$$  \hspace{1cm} (3)

where $S(\ast)$ is the segmentation branch and $F^0$ denotes the finial feature maps of the coarse network. After the first refinement, semantic maps will be updated as the segmentation $S_{RN}^1$ (the superscript indicates the number of iterations) corresponding to the fine predictions for next iteration to guide inpainting. The same rules are followed in subsequent semantic maps updates. To suit our task and make effective use of the semantic priors and generative facial priors, we put forward the WPGM layer to adaptively modulate feature maps.

**C. WPGM LAYER**
We devise a WPGM layer (called the Weighted Prior-Guidance Modulation layer), its process shown in Fig 3. This spatial feature transformation layer generates learnable modulation parameter pair based on semantic information and generative facial prior, and then does wise-element multiplication and addition. In order to facilitate later modulation, we combine semantic maps with generative facial priors. Then, they are convolved to generate intermediate condition from prior information and then produce the modulation parameter pair with the same dimension as $F_{RDEC}$. The feature maps in the decoder will be influenced by an affine
transformation spatially. Our goal is to derive the result close to ground truth as much as possible. For the boundary of holes, it can be easily inferred from contextual features. But for the centre, it lacks enough known information to inpaint this area. Therefore, the parameters need to be weighted and then the features of $F_{SGAN}$ play a larger role in the centre of holes.

According to the mask of input images, we devise a method of exponential decay to derive the weight map. In an adjacent area centred on the target pixel, count the number of known pixels (i.e., zero in the mask), and divide by the total number of pixels in this $k \times k$ area. Therefore, the less known pixels, the larger the value of the weight, which is consistent with demands.

\[
  w_{x,y} = \alpha^{-\left(\frac{n}{w_{x,y}}\right)}, \quad (x, y) \in M \left(\frac{-\frac{k}{2} - \frac{1}{2}}{2}, \frac{\frac{1}{2}}{2}\right) \rightarrow (x+\frac{k}{2}, y+\frac{1}{2}) \quad (4)
\]

\[
  \Psi^l = (W \times F^l_{SGAN}) \oplus S^l \quad (5)
\]

where $\alpha > 1$, and $w_{x,y}$ denotes the weight value at location $(x, y)$. $M$ denotes the mask and $l$ denotes the scale. $\oplus$ refers to channel addition and $\times$ means that weighted map $W$ multiplies with $F_{SGAN}$ in every channel. Let $F_{RDEC} \in \mathbb{R}^{H \times W \times C}$ as the input feature maps.

\[
  F_{out} = concat\left( y^l(\Psi)^l_{RDEC} - \frac{\mu^l}{\sigma^l} + \beta^l(\Psi), F^l_{RDEC}\right) \quad (6)
\]

where the $y^l(\Psi)$ and $\beta^l(\Psi)$ represent the learned modulation parameter pair. $F^l_{RDEC}$ denotes the feature maps of $l^{th}$ layer in the decoder of the refinement network. The $\mu^l$ and $\sigma^l$ denote the mean value and standard deviation of the activations, respectively. $concat$ refers to concatenation operation. After modulation operation, the output $F^{l+1}_{out}$ will be passed through vanilla convolution to further infer features.

\[
  F^{l+1}_{RDEC} = Conv(F^{l}_{out}) \quad (7)
\]

where $Conv$ denotes vanilla convolution. The WPGM layer leverages two kinds of prior information well for spatial feature modulation, achieving a good balance between richness and fidelity. In addition, it reduces complexity because it requires fewer channels for modulation.

**D. SEMANTIC EVALUATION**

We also adopt the iterative network structure, as in the case of [25], [30], but the difference is that we make evaluation about the segmentation of inpainting results to update masks for next iteration instead of generating confidence maps. When the max probability of semantic pixels in $c$ channels is lower than threshold, we set it as one to update the mask. Therefore, one in the mask displays the location of pixels with ambiguous semantics.

\[
  M^i(x, y) = \begin{cases} 
  1, & P^i_{Rmax} < \tau \\
  0, & \text{otherwise}
  \end{cases} \quad (8)
\]

where $P^i_{Rmax}$ represents the max confidence scores of the segmentation confidence map of the inpainting result in $i^{th}$ iteration and $\tau$ is threshold. $M(x, y)$ represent the pixel value at $(x, y)$ in the mask. In the new mask, white regions are where probably needs to be re-inpainted. Thus, in the next iteration, our model will reason features of holes again, according to well-inpainted regions and known pixels. We iterate third during training and fix the number of iteration while testing. It is worth noting that the next iteration starts from the refinement network. There is no need to make a coarse prediction again, because the model has derived approximate content of damaged images through previous iterations. At this moment, the inpainting result of last iteration and the updated mask will be considered as the new input and use the final segmentation maps of last iteration as a guide for inpainting for the next iteration. Other parts follow the first round of operations and parameters are shared for each round of iteration.

**Algorithm 1** The GSGIFI Network

**Input:** Input image $Im_{gin}$. Input mask $Mask$

**Output:** Output reconstructed image $I_{out}$

**Denotation:** “$\odot$” as the multiplication operation

**Begin Algorithm**

1. Get coarse prediction $I_{CN} = PartialUNet(Im_{gin}, Mask)$
2. Refer to (3) to get segmentation $S_{CN} = S(F^3)$
3. Let $S_p = S_{CN}$, $I_{in} = concat(I_{CN}, Mask)$
4. for $i \in \{1, \ldots, q\}$ do
5. Refer to (1) to generate the latent code $z$
6. Refer to (2) to get $F_{SGAN}$
7. $F^4_{RDEC} = Decoder(F_{latent}F_{SGAN}, S_p)$
8. Get fine prediction $I_{RN} = N(F^4_{RDEC})$
9. Get segmentation $S_{RN} = S(F^4_{RDEC})$
10. Refer to (8) to update mask $M$
11. Initialize input $I_{in} = concat(I_{RN} \odot M, M)$, $S_p = S_{RN}$
12. $i = i + 1$
13. end
14. return $I_{out} = I_{RN}$

**E. OBJECTIVE**

1) FACIAL FEATURE SELF-SYMMETRY LOSS

Inspired by face symmetry, we further devise a loss to generate the symmetrical outputs by minimizing the distance in feature space rather than pixel space. We argue that the attributes (i.e., gender, shape, colour, style) of the left face should be similar to the right face. But for most of existing algorithms, they may produce asymmetric inpainting face images, e.g., male right eye with female left eye. Therefore, we devise a self-symmetry loss function to constrain left and right face to be coordinated in the case of frontal face and mild side face. Considering that when there is a large side face situation, we only constrain the symmetry of the left and right face to be coordinated. First, we get the ROI face of the repaired results and split it into left and right face. They are resized and input into face extractor [42] to extract features.
separately and we will calculate the distance between them.

\[
L_{sym} = \sum_{c} \frac{1}{H \times W} \sum_{H} \sum_{W} \| \phi(\text{ROI}_{left}) - \phi(\text{ROI}_{right}) \| \tag{9}
\]

where \( \phi(*) \) is to extract features of ROI region separately and \( \text{ROI}_{left}, \text{ROI}_{right} \) are the left and right of ROI face respectively. \( H, W, C \) are the height and width and channel of the final output. We minimize the gap between right and left face in the feature space where features are higher-dimensional and abstract, so that we can derive a face with symmetrical style. Besides, L1 distance is used as reconstruction loss to constrain the inpainting results.

\[
L_{re} = \lambda_{sym} L_{sym} + \lambda_1 \| I_{out} - I_{gt} \|_1 \tag{10}
\]

where \( \lambda_{sym}, \lambda_1 \) are the weight of L_{sym}, L1. Besides, \( I_{gt} \) is the ground truth and \( I_{out} \) is the generated result.

2) ADVERSARIAL LOSS

Similar to [34], we employ an adversarial loss to distinguish synthesized face images from real images to produce more realistic images.

\[
L_{adv} = -\lambda_{adv} E_{out}\text{softplus}(D(I_{out})) \tag{11}
\]

where \( \lambda_{adv} \) is the weight of L_{adv} and \( D \) represents the discriminator. \( L_{adv} \) acts as a supervisor to promote generated images to be realistic, avoiding blurring and producing sound effects visually.

3) MULTI-ITERATION CROSS-ENTROPY LOSS

In order to make the generated semantic maps close to ground truth, the semantic maps \( S^i_{RN} \) obtained at the end of \( i^{th} \) iteration are penalized at each position. The formula is as follows:

\[
L_{cross}(S, S^i_{RN}) = -\lambda_{cross} \sum_i \sum_m S_m \log S^i_{RN} \tag{12}
\]

where \( m \) represents each pixel at the semantic map of ground truth and \( \lambda_{cross} \) is the weight of L_{cross}.

4) IDENTITY PRESERVING LOSS

Considering that identity information plays an important role in developing “recognition via inpainting”, we use identity preserving loss like [43] to narrow the distance of identity information between completed face and ground truth. Due to the excellent ability of pre-trained light-CNN [44] in capturing face identity information, we use it to extract features and calculate the distance of its output. Specifically, we define identity preserving losses as follow:

\[
L_{id} = \lambda_{id} \frac{1}{N} \sum_{n=1}^{N} |\Gamma(I_{gt}) - \Gamma(I_{out})| \tag{13}
\]

where \( \Gamma(*) \) denotes identity recognition network and \( \lambda_{id} \) represents the weight of this loss. \( N \) denotes the dimension and \( I_{gt}, I_{out} \) denote ground truth and final output of our model, respectively. Finally, the total loss function is organized as:

\[
L_{total} = L_{re} + L_{adv} + L_{cross} + L_{id} \tag{14}
\]
FIGURE 5. Examples cases of visual comparison on the CelebA-HQ datasets.

TABLE 1. Hyper-parameter settings. \(\lambda\) represents the weight of loss. \(\beta\) is the hyperparameter of adam.

| Hyper-parameter | Value | Hyper-parameter | Value |
|-----------------|-------|-----------------|-------|
| \(\lambda_{sym}\) | 0.02  | \(\lambda_1\)   | 1     |
| \(\lambda_{cross}\) | 6     | \(\beta_1\)    | 0.9   |
| \(\lambda_{adv}\)   | 1e-3  | \(\beta_2\)    | 0.99  |
| \(\lambda_{ld}\)    | 3\times1e-2 | Learning rate | 1e-4  |

from the CelebA dataset. When selecting images out as CelebA-HQ, delete them in the CelebA dataset. There are 172,599 images left in the CelebA dataset. CelebA-HQ dataset has semantic labels with 19 categories. Except for facial components, there are 8 classes such as neck, eyeglass, earring, necklace, and cloth. We use CelebA-HQ dataset to train model and evaluate the effect of the inpainting approach and leverage the images of CelebA dataset (not include images in CelebA-HQ) to verify recognition via inpainting. We take about 2,000 images from CelebA-HQ dataset for inpainting test. The training dataset is cropped with centre square masks and random masks. The masks used for training come from a random mask dataset and other block masks that are randomly generated. In the verification experiments of recognition via generation, we use 156,000 images of CelebA dataset, 150,000 images for training, 3,000 images for validation and 3,000 images for test.

LaPa dataset is a large challenging public dataset newly released in 2020, covering huge variations in expression, occlusion and pose. It contains over 22,000 images with 11 semantic categories less than CelebA-HQ dataset and we use 18,176 images for training, 2,000 images for validation and 2,000 images for test.

B. COMPARISON MODEL
We compare our model with several start-of-the-art approaches on CelebA-HQ and LaPa dataset. These inpainting models are: 1) PC[1]: partial convolution is proposed to complete the inpainting task. 2) LaFIn[6]: the inpainting method based on facial landmarks. 3) RFR[25]: the inpainting method of recurrent feature reasoning. 4) GC[20]: gated convolution is proposed for inpainting. 5) SF[14]: two-stage model based on structure generator and texture generator. We directly use the officially realised pretrained model of GC. We used their official codes for the rest of the models, and they are fine-tuned on LaPa dataset.

C. COMPARISON WITH STATE-OF-THE-ARTS
1) QUALITATIVE COMPARISON
Fig. 5 and Fig. 6 are visual comparison. A qualitative comparison on the CelebA-HQ dataset is shown in Fig. 5. Although GC and PC can complete the holes, the contents are not very realistic and detailed. The details of the eyes repaired by the PC are not very clear. In Fig. 5, generated eyes and eyebrows by GC are very light in color. The texture of RFR is blurrier. However, our model generates more detailed and realistic eyes with clear boundary. Fig.6 shows the quantitative comparison on LaPa dataset. LaPa dataset is more challenging for inpainting and has many variations in illumination, pose, expression, etc. As shown in Fig. 6, PC struggles to fill holes with correct content for images with occlusion and pose. RFR also performs poorly on LaPa dataset, and its results are prone to distortion. The failure of LaFIn to repair the nose probably due to predicting wrong landmarks as shown in (d) in Fig. 6. Similarly, SF does not have very clean structure. In contrast to these methods, our model can better handle real occlusion and restore the natural illustration. While missing area is too large as shown in second row in Fig.5, it is almost impossible to restore face perception similar to ground truth, even for humans. All in all, our model outperforms above models, especially in detail and symmetrical attributes.

2) QUANTITATIVE COMPARISON
We use PSNR, SSIM, FID and L1 loss as metrics to evaluate the results. The higher the value of PSNR and SSIM, the better the image quality. FID and L1 loss are the opposite. FID

VOLUME 10, 2022
represents fidelity of images. Table 2 provides comparison results on CelebA-HQ and LaPa datasets. It is demonstrated that LaFin has great performance over RFR and PC on all of metrics with FID of 3.512, SSIM of 0.912. SF takes into account texture and structure information and it also has better scores. RFR gets lower metric scores with SSIM of 0.773, which is likely caused by its simple feature reasoning module. It can be observed that scores of our model are superior to LaFin (PSNR: 27.12 vs 26.25) and SF (FID: 3.24 vs 5.03). At the same time, it is easy to find out that performance of all models on CelebA-HQ is better than on LaPa. This is because the former dataset is almost frontal face with simple background and has less variation in occlusion, pose and ethnicity. On this challenging dataset, our model still has the superiority over other methods with PSNR of 26.17 and L1 error of 0.324. Above all, the proposed model performs better than other methods on all metrics.

### D. RECOGNITION VIA INPAINTING

During the epidemic, the missing feature caused by wearing a mask makes face recognition more challenging. We hope to test whether a finer inpainting model can improve recognition accuracy under larger occlusion. Previously, Mathai et al. [48] conducted similar experiments under occlusion by hands, sunglasses, etc., using a simple inpainting model. It is worth noting that this experiment is performed on a single image, instead of a real-time scene. Although there are lots of state-of-the-art inpainting approaches, they focus more on image quality and results are not shown to be suitable for recognition tasks. To verify capabilities of restoring face perception of our model, we pass inpainting results through Light-CNN [44] to extract features and measure the Rank-1 recognition accuracy with the cosine-distance metric.

We verify on 156,000 images from CelebA datasets and split them into training, validation, and test datasets as in section A of part IV mentioned above. All images on datasets are occluded by the shape of real objects through pre-processing. Recognition models (Light-CNN [44]) and face extractors (i.e., ResNet-50 [42]) are trained on training dataset. Our inpainting model has been trained on CelebA-HQ and is directly used for this experiment. For other inpainting models, we use their pretrained model on CelebA dataset. During test, one is to pass damaged images through the inpainting network to a trained Light-CNN to verify whether inpainting can improve recognition accuracy. Another is to directly use backbones to extract face feature and get classification probability. In addition, we also input our inpainting results into ResNet-50 to see how performance is. Table 3 provides the recognition accuracy of the two ideas. We can see that the effect of results without inpainting is uncompetitive with recognition via inpainting. The performance of Light-CNN is extremely worse as the holes becoming large. The effect of the way to cascade inpainting model with the face feature extractor is much better than single backbone in the mild occlusion. Moreover, our model
E. MODEL EFFICIENCY

To comprehensively evaluate our model, we compare our model with other state-of-the-arts in complexity and performance, as shown in Table 4. FLOPs metric is used to measure the model complexity by measuring the float point operations of a model under 256 × 256 resolution. It shows that our model has best results compared with other models. RFR requires only 0.15s and smallest parameter amount but their results are not satisfying among comparison models. SF is more complex as well as has higher PSNR and SSIM values. LaFIn has larger model size of 97M and FLOPs of 102.9G, with better results. In contrast with them, our GSGIFI network has best SSIM and PSNR scores with smaller sized model of 89M and less time of 0.207s. The FLOPs and parameter amount of our model are larger than RFR, but the performance is much better (23.53dB vs 25.21dB). Taken together, we conclude that our model well balances the complexity, time consumption and performance, which also highlights the efficiency of our model.

F. ABLATION STUDIES

1) EFFECTIVENESS OF THE WPGM LAYER

In the proposed network, one of the core components of our approach, the WPGM layer fuses two kinds of prior information to guide the fine inpainting. In order to study the effectiveness of WPGM layer and prior information, we study the ablation of three variants: 1) WPGM layer without weight maps (w/o weight); 2) WPGM layer without semantic maps (w/o sem); 3) WPGM layer without generative facial priors (w/o gen). As shown in the third column of Fig. 7, without the weight map, we can see that the inpainting results are more different from the ground truth, which is influenced by the generative priors. Table 5 shows little difference. The

TABLE 3. Rank-1 recognition rates (%) across views under different ratio of occlusion.

| Method                | 10%-25% | 25%-40% | 40%-55% |
|-----------------------|---------|---------|---------|
| ResNet-50             | 77.23   | 51.71   | 43.32   |
| Light-CNN             | 71.6    | 45.18   | 19.87   |
| LaFIn+Light-CNN       | 80.62   | 63.23   | 52.04   |
| Ours+Light-CNN        | 86.91   | 78.27   | 61.98   |
| Ours+ResNet-50        | **89.47** | **81.58** | **65.41** |

TABLE 4. Comparison of model complexity and time consumption and effectiveness of different models.

| Method                | RFR | SF | LaFIn | Ours |
|-----------------------|-----|----|-------|------|
| Model size(M)         | 34  | 84 | 97    | 89   |
| Time cost(s)          | 0.153 | 0.209 | 0.224 | 0.207 |
| FLOPs(G)              | 40.5 | 89.4 | 102.9 | 78.6 |
| PSNR                  | 23.53 | 24.76 | 24.32 | 25.21 |
| SSIM                  | 0.727 | 0.769 | 0.775 | 0.798 |

performs better than LaFIn, which also shows that our model performs well in face perception restoration. Table 3 also shows that our inpainting model works better with ResNet-50. However, as the holes getting larger, the effect is limited. In this case, it is hard for models to restore face perception similar to ground truth.
weight map has little impact on the image quality but on face perception, which is why metrics are unable to tell the distinction.

a: SEMANTIC PRIOR
As shown in the second column (b) in Fig. 7, the model fails to repair the right eye in the first row. Look at (b) in the third row, it’s a bad result that the right face is extremely unmatched with left face. Without semantic maps for guidance, the model produces terrible structure and content. From Table 5, we can find out that the evaluation scores are inferior to the proposed model (with all). It also indicates that semantic priors have a large impact on the inpainting results than other factors. We further visualize semantic maps to validate its effectiveness. As can be seen in Fig. 8, the segmentation of coarse prediction in the first iteration may be not very accurate at the edge of holes. But for most content pixels, the label is right which can guide the refinement and the semantic maps are updated at the end of one iteration. Besides, segmentation by our method is very close to ground truth. Above all, it indicates that our semantic priors are effective, demonstrating the reliability of our semantic-guided method.

b: GENERATIVE FACIAL PRIOR
From the second column in Fig. 7, the inpainting images are not very clear and have inconsistent attributes. The repaired eyes in the first and second rows in (c) are blurry and uncoordinated. Generative facial priors contain rich textures and style information so that the model is able to complete content. Table 5 also proves that generative facial priors can improve the image quality. Without these priors, the model tends to more likely produce unnatural facial attributes and unreal textures.

2) EFFECTIVENESS OF FACIAL FEATURE SYMMETRY LOSS
The loss function can constrain inpainting face images to be symmetrical, which makes results more in line with human vision. As shown in Fig. 7, we can easily find that results without loss (d) are inferior to (f) which is more symmetrical, i.e., similar left and right eyes and eyebrows. These changes may not be visible in the evaluation metrics as Table 5 shows, because they are not enough to evaluate these perceived features which is easy to human perceptual system.

3) ITERATION COMPARISON
Fig. 9 shows the $L_1$ error on validation set of LaPa dataset with free-form masks, under different numbers of iteration given specific occlusion ratios. While 25%-40% percent of areas are occluded, iterated three times for the best results during training. We observe that the results are getting better with the increasing iterations, which meets our expectations that the holes are progressively inpainted well during iteration. After reaching the best effect, more iterations will not continue to improve the performance. As the occlusion area increases, the number of iterations required also increases. When the occlusion area is too large, the $L_1$ error of the real image will naturally be higher. In general, under different occlusion rates, three iterations can achieve good performance, so we set iteration number as 3 for LaPa dataset during test.

G. LIMITATION
Although our method has achieved good results in many situations, there are still extreme cases that is difficult for the model. Here we provide failure cases as shown in Fig. 10.
In the first row, we can find that the inpainting result hardly matches the human visual perception. Long nose and lip-sticked mouth are not compatible with the child’s identity. The training dataset is almost entirely of adults so that the network is insensitive to images of children, and the face image tends to be inpainted with features of adult women. At the same time, the original image has a large expression and opening the mouth makes the face longer, which also increases the difficulty of inpainting. In the second row, the input image is a largely occluded profile face and its inpainting result is distorted and blurred. It is difficult to repair the full profile face, and the boundary between the background and the face is also difficult to distinguish, which causes the structure to be chaotic and blurred. But for most of the micro-profile faces, our model can perform very well.

V. CONCLUSION AND FUTURE WORK

We have proposed a novel iterative inpainting model with semantic evaluation for generating coordinated and symmetrical face images with reasonable content and edges. Our approach integrates generative facial prior and semantic information into refinement to modulate feature maps by normalization layer. To protect original information, we devise an exponential decay weight map to mask the generative facial priors. With above two kinds of priors, the model can supplement rich facial information when reasoning the content of missing holes. We further devise a self-symmetry loss to constrain symmetry of face completion. Quantities of experiments prove our design is more stable than the two-stage architecture and is superior to the state-of-the-art methods in detailed textures and consistent attributes. Moreover, experiments show that our model can help improve the accuracy of face recognition under mask occlusion. Our model can also be adapted for natural image inpainting. For example, remove generative face priors and iteratively repair a natural image based on the predicted semantic map. For another example, it is more likely to recover finer structures by replacing generative face priors with generated scene priors included in pretrained GANs.

AUTHOR CONTRIBUTIONS

Xin-Yu Zhang conceived the algorithm of the paper and designed the experiments; Kai Xie reviewed the paper; Mei-Ran Li conducted comparative experiments; Chang Wen collected data and conducted ablation studies; and Jian-Biao He checked spelling and grammar and made suggestions.

REFERENCES

[1] G. Liu, F. A. Reda, K. J. Shih, T. C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolution," in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 85–100.
[2] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," ACM Trans. Graph., vol. 36, no. 4, pp. 1–14, Jul. 2017.
[3] 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.
[4] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4170–4179.
[5] Y. Zhang, F. Ding, S. Kwong, and G. Zhu, "Feature pyramid network for diffusion-based image inpainting detection," Inf. Sci., vol. 572, pp. 29–42, Sep. 2021.
[6] Y. Yang, X. Guo, J. Ma, L. Ma, and H. Ling, "LaFIn: Generative landmark guided face inpainting," 2019, arXiv:1911.11394.
[7] Y. Song, C. Yang, Y. Shen, P. Wang, Q. Huang, and C.-C. Jay Kuo, "SPG-Net: Segmentation prediction and guidance network for image inpainting," 2018, arXiv:1805.03566.
[8] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edge-Connect: Structure guided image inpainting using edge prediction," in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 3265–3274.
[9] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5833–5841.
[10] J. Li, F. He, L. Zhang, B. Du, and D. Tao, "Progressive reconstruction of visual structure for image inpainting," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 5961–5970.
[11] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 2332–2341.
[12] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 606–615.
[13] A. Dundar, K. Sapra, G. Liu, A. Tao, and B. Catanzaro, "Panoptic-based image synthesis," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 8067–8076.
[14] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "StructureFlow: Image inpainting via structure-aware appearance flow," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 181–190.
[15] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," in Proc. Eur. Conf. Comput. Vis. (ECCV), Glasgow, U.K., Aug. 2020, pp. 725–741.
[16] D. Ping, S. Ram, and J. I. Rodriguez, "Image inpainting using nonlocal texture matching and nonlinear filtering," IEEE Trans. Image Process., vol. 28, no. 4, pp. 1705–1719, Apr. 2019.
[17] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 9168–9178.
[18] B. Mandal, A. Okekewu, and Y. Theis, "Masked face recognition using ResNet-50," 2021, arXiv:2104.08997.
[19] S. M. N. Uddin and Y. J. Jung, "Global and local attention-based free-form image inpainting," Sensors, vol. 20, no. 11, p. 3204, Jun. 2020.
[20] 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. 4470–4479.
[21] S. Yang, Z. Wang, J. Liu, and Z. Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," in Proc. Eur. Conf. Comput. Vis. (ECCV), Aug. 2020, pp. 601–617.
[22] L. Song, J. Cao, L. Song, Y. Hu, and R. He, "Geometry-aware face completion and editing," in Proc. AAAI Conf. Artif. Intell., 2019, pp. 2506–2513.
[23] T. Wang, H. Ouyang, and Q. Chen, "Image inpainting with external-internal learning and monochromic bottleneck," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 5120–5129.
[24] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6882–6890.
