Collaborative Contrastive Learning-Based Generative Model for Image Inpainting

YONGQIANG DU1, HAORAN LIU1, AND SONGNAN CHEN2
1School of Information Engineering, Xinyang Agriculture and Forestry University, Xinyang 464000, China
2School of Mathematics and Computer, Wuhan Polytechnic University, Wuhan 430048, China
Corresponding author: Songnan Chen (chensongnan@whpu.edu.cn)

This work was supported in part by the Science and Technology Department of Henan Province under Grant 182102110160 and Grant 222102110189, and in part by the Young Teachers Found of Xinyang Agriculture and Forestry University under Grant 201701013.

ABSTRACT The critical challenge of image inpainting is to infer reasonable semantics and textures for a corrupted image. Typical methods for image inpainting are built upon some prior knowledge to synthesize the complete image. One potential limitation is that those methods often remain undesired blurriness or semantic mistakes in the synthesized image while handling images with large corrupted areas. In this paper, we propose a Collaborative Contrastive Learning-based Generative Model (C2LGM), which learns the content consistency in the same image to ensure that the inferred content of corrupted areas is reasonable compared to the known content by pixel-level reconstruction and high-level semantic reasoning. C2LGM leverages the encoder-decoder based framework to directly learn the mapping from the corrupted image to the intact image and perform the pixel-level reconstruction. To perform semantic reasoning, our C2LGM introduces a Collaborative Contrastive Learning (C2L) mechanism that learns high-level semantic consistency between inferred and known content. Specifically, C2L mechanism introduces the high-frequency edge maps to participate in the process of typical contrastive learning and enables the deep model to ensure the semantic reasonableness between high-frequency structures and pixel-level content by pushing the representations of inferred content and known content close and keeping unrelated semantic content away in the latent feature space. Moreover, C2LGM also directly absorbs the prior knowledge of structural information from the proposed structural spatial attention module, and leverages the texture distribution sampling to improve the quality of synthesized content. As a result, our C2LGM achieves a 0.42 dB improvement over competing methods in terms of the PSNR metric while coping with a 40~50% corruption ratio in the Places2 dataset. Extensive experiments on three benchmark datasets, including Paris Street View, CelebA-HQ, and Places2, demonstrate the advantages of our proposed C2LGM over other state-of-the-art methods for image inpainting both qualitatively and quantitatively.

INDEX TERMS Image inpainting, semantic reasoning, contrastive learning, generative model.

I. INTRODUCTION Image inpainting, which aims to fill missing regions with plausible content, is a fundamental problem among image restoration tasks [1], [2] and is prominently developed by the wide applications of deep convolution neural networks (CNNs). Typical CNN-based approaches [3], [4] for image inpainting employ an encoder-decoder structure to predict the content of corrupted regions, and they can perform well in recovering minor defects. However, there are two critical issues for existing CNN-based methods that need to be addressed. 1) Since the corruption of input images simultaneously loses structure and texture, it is hard to directly leverage CNN models to synthesize exquisite details from limited prior knowledge, especially handling large continuous corrupted regions; 2) Typical CNN-based methods for image inpainting model the mapping from the corrupted image to the intact image, which inevitably leads to inpainting...
implausible semantics for the lack of explicitly high-level semantic learning. Therefore, it is of great importance to developing an image inpainting framework that explicitly learns to synthesize reasonable content for the corrupted regions from practical high-level guidance.

In recent years, many methods have been proposed to alleviate the aforementioned problems. A prominent research line that improves the performance of image inpainting [5], [6], [7], [8], [9], [10], [11] employs specific prior knowledge to facilitate content inference, which follows two typical ways. A straightforward way of prior-based methods [5], [6], [7] leverages mask as prior knowledge to guide the content reconstruction of masked regions. Whilst such inpainting paradigm is validated its effectiveness, especially for irregular image inpainting, it often fails to cope with large and continuous corruption. The other typical way [8], [9], [10], [11] of prior-based methods for image inpainting predicts the structural information and then leverages it as guidance to facilitate inferring the corrupted content. Nevertheless, the semantics in the reconstructed image is usually unreasonable due to the errors of predicted structural maps, and it is difficult for deep models to learn similar semantics from training samples implicitly.

Another research line [12], [13], [14], [15], [16], [17], [18] of image inpainting is to learn the semantic consistency between filled regions and known regions. Inspired by traditional patch-based methods [19], [20], [21], some research [12], [13] computes the content similarity between image patches and selects similar patches as replacement cues to infer desired content. However, these methods bring limited improvement mainly due to the lack of general semantic understanding. With the promising effects of the attention mechanism [22] for semantic understanding, some methods [14], [15], [16], [17], [18] leverage it to learn semantic consistency between the inferred and known regions. While they learn the semantic consistency of various content, undesired blurriness or implausible structure still exists when the corruption area is large. Even if the inferred content is semantically identical to the known region, its texture or structure may still be diverse. Furthermore, as illustrated in Figure 1, current methods for image inpainting suffer from severe performance degradation and synthesize implausible content as the increasing corruption ratio. Thus, it remains a grand challenge to cope with large corrupted areas for image inpainting.

Although current methods often cannot work as expected, some heuristic research provides significant insights for our research. Some recent research builds a probabilistic model to estimate the distribution of corrupted content as priors [23], [24]. They employ variation-autoencoder(VAE) [25] as the main framework and optimize parameters by the Kullback-Leibler divergence. In addition, recent generative models [26], [27], [28] possess the ability of generative representation of global distribution for training samples and provide appropriate content for the specific corrupted image.

In this paper, we propose the Collaborative Contrastive Learning-based Generative Model (C2LGM) for image inpainting, which introduces a novel contrastive learning strategy that aggregates multi-view information, such as high-frequency edge or pixel-level content, from the known regions to collaboratively keep high-level semantic consistency of inpainted image. In contrast to previous methods using contrastive learning, our C2LGM is a novel framework that simultaneously employs typical pixel-level reconstruction from the outside to the inside and performs semantic inference by contrastive learning, which pushes similar semantics close and keeps unrelated semantics away. Thus, primary pixel-level reconstruction can be strengthened by the guidance of semantic priors learned by contrastive learning, which facilitates inpainting plausible content even in a large corruption ratio. Moreover, the collaborative contrastive learning strategy of C2LGM adopts a multi-view manner that establishes contrastive pairs from pixel-level patches and high-frequency patches to emphasize the consistency between semantic structures. As a result, C2LGM is able to inpaint reasonable content in large corrupted areas and makes it as realistic as natural images. The main contributions in this paper can be summarized as follows:

- We propose a novel image inpainting framework that fills reasonable content for large corrupted regions by collaborating pixel-level reconstruction and high-level semantic reasoning. By learning prior knowledge from high-frequency edge inference, our C2LGM is able to synthesize content as realistic as natural images, especially for large corrupted regions.
- To keep high-level semantic consistency of inferred and known content and synthesize reasonable high-frequency structures, C2LGM introduces a collaborative contrastive learning strategy. Besides, we propose a texture distribution learning method to ensure pixel-level texture consistency, which facilitates inpainting exquisite details for large corrupted regions.
- Extensive experiments on three benchmarks of image inpainting, including Paris Street view, CelebA-HQ, and Places2, demonstrate both the quantitative and qualitative advantages of our proposed C2LGM.

Considering the randomly selected example presented in Figure 1, our C2LGM employs two collaborative decoders for both image and edge reconstruction. Differing from the common way that only leverages an encoder-decoder based framework [29] to inpaint the corrupted regions from the outside to the inside [5], [6], [7], our C2LGM leverages a shared encoder to perform contrastive learning, and then provides reliable semantic consistency between the inferred and known regions. Current state-of-the-art methods for image inpainting become gradually incapable with the increasing corruption ratio. Instead, our C2LGM can infer according to much limited known information with the high-level guidance by collaborative contrastive learning (see Figure 3), and fills in more reasonable structure and realistic details than other state-of-the-art methods (see Figure 4).
II. RELATED WORK

Over the past decades, there have been several researches on image inpainting. In this section, we look back to the most related methods, which can be divided into two categories: non-learning based methods and deep learning-based methods.

A. NON-LEARNING BASED METHODS FOR IMAGE INPAINTING

As a traditional solution to image inpainting, non-learning based methods mainly consist of two categories, diffusion-based methods and exemplar-based methods. The diffusion-based methods [30], [31], [32] diffuse existing pixels from the known regions to the corruption regions. By regarding texture consistency between corrupted and known content, diffusion-based methods are able to fill plausible content for images with tiny corruptions. Nevertheless, such methods often synthesize undesired blurriness and unreasonable semantics.

To relieve above problems, the exemplar-based methods [19], [20], [21], [33] select the most similar patch from known regions or other images to replace the corrupted regions. Most methods seek to design algorithms to select the optimal patch. Criminisi et al. [33] introduce a method to calculate the confidence values of corrupted regions for inpainting orders. Barnes et al. [19] propose the PatchMatch algorithm that selects the optimal patch from nearest neighbor patches in a quick manner. Exemplar-based methods are able to inpaint reasonable content for images with repeated textures. In sum, most non-learning based methods for image inpainting often generate unreasonable semantics and undesired artifacts (blurriness, noise, etc.) due to the lack of semantic understanding. Thus, such methods are not able to satisfy the needs of inpainting reasonable content for diverse real-life scenes in present image inpainting.

B. DEEP LEARNING-BASED METHODS FOR IMAGE INPAINTING

1) MASK-GUIDED METHODS FOR IMAGE INPAINTING

The excellent capability of feature learning by deep convolutional neural networks (CNNs) brings prominent improvement in inpainting performance. Pathak et al. [4] first design a deep CNN model for image inpainting, and employ a context-encoder to infer corrupted content in the latent feature space. Later, many CNN-based methods for image inpainting employ mask as guidance to infer the corrupted regions from the outside to the inside. For inpainting irregular corruption, Liu et al. [5] propose the partial convolution to distinguish the inferred content and known content, and avoid appearing undesired noise by updating the mask in the feature space. To improve the effectiveness of mask guidance, Yu et al. introduce the gated convolution [6] to learn and update soft masks in the latent feature space for weighting the confidence of inferred pixels. Later, Zheng et al. [34] propose a masked spatial-channel attention mechanism and patch-based self-supervision strategy for image synthesis. Chen et al. [35] introduce a hierarchy framework to aggregate multi-model in formation and improve the quality of synthesized image. To make deep models directly distinguish the inferred and known content, Yu et al. [36] propose the region normalization. To iteratively infer the corrupted content, Li et al. [7] design a framework to progressively synthesize the missing content with the mask guidance by partial convolution. The mask-guided methods for image inpainting are able to handle irregular corruption, but often fail to synthesize reasonable content for large continuous corruption.

2) STRUCTURAL KNOWLEDGE-GUIDED INPAINTING METHODS

Structural priors are crucial guidance for the reconstruction of corruption content. Liu et al. [16] learn coherent semantic relevance from known regions by the attention mechanism [22]. Nazeri et al. [8] propose a two-stage framework that first predict the edge map of complete image, and then reuse it as structural guidance to reconstruct the intact image. Xiong et al. [39] employ the foreground mask as guidance to keep the completeness of salient objects in corrupted images. Yang et al. [10] combine the structural information and pixel-level information to improve the stability of semantic synthesis. Considering both structure and texture reconstruction, Liu et al. [37] design a mutual encoder-decoder based framework to improve the quality of synthesized content. Roy et al. [40] present an image inpainting method using frequency-domain information as structural priors to reconstruct the intact image. Feng et al. [41] introduce a deep-masking mechanism to infer corrupted areas from the known content in a coarse-to-fine manner. Bo et al. [42] propose a two-stage image inpainting model composed of structure generation network and texture generation network. Ji et al. [43] propose a nested transformer architecture for image synthesis. Similarly, Guo et al. [38] propose a two-stream network dividing image inpainting task into structure-constrained texture synthesis and texture-guided structure reconstruction. The structural knowledge-guided methods provide high-frequency cues for reconstructing the complete image, but the error of structural information still leads to unreasonable semantics of corrupted regions.

3) GENERATIVE PRIOR-GUIDED METHODS FOR IMAGE INPAINTING

With the fast development of generative models [25], [44], some recent methods for image inpainting attempt to learn generative prior for performance improvement. Li et al. [45] design a generative adversarial net (GAN) based model to force the distribution of inferred content similar to the known content. Zheng et al. [23] propose a Variation Auto-Encoder (VAE) based framework to synthesize pluralistic content for the corruption regions. Inspired by exemplar-based methods, Zhao et al. [24] leverage random samples as an exemplar to inpainting diverse content by the
FIGURE 1. Main framework of our proposed C2LGM. The input image and mask are first sent to the encoder E to extract features, and two parallel decoders, $D^p$ and $D^e$, are employed to predict the complete image and edge map. Finally, the collaborative contrastive learning approach facilitates the clustering of similar semantics in the latent space. Consequently, our model is able to inpaint more reasonable content than other state-of-the-art methods, including EdgeConnect [8], MEDFE [37], and CTSDG [38].

VAE model. Even though excellent results are conducted in recent NTIRE image inpainting challenges [46] in terms of unsupervised and label-guided image inpainting, a large corrupted region is often intractable for many methods for the image inpainting task. In summary, present generative methods for image inpainting are hard to infer reasonable semantics and often synthesize implausible content for large corruption. Besides, semantic reasoning by contrastive learning is not fully explored in the image inpainting task, and thus we establish contrastive pairs from both the pixel-level information and high-frequency edge information to facilitate the semantic understanding of similar content.

III. METHOD

Image inpainting aims to infer the content of the corrupted regions, which is required to be consistent with known content on semantics and sharpness. To this end, we propose the Collaborative Contrastive Learning-based Generative Model (C2LGM), which employs contrastive learning to keep consistency between the inferred and known content. Moreover, C2LGM employs two respective decoders to reconstruct the complete image, and they leverage the structural spatial attention mechanism to facilitate the reconstruction of high-frequency structures.

A. MAIN FRAMEWORK OF C2LGM

The C2LGM is a cooperative framework that infers both the edge (high-frequency) map and intact image, and the detailed framework of C2LGM is elaborated in Table 1. As illustrated in Figure 1 and Figure 3, the encoder of C2LGM is leveraged not only to capture features from input corrupted images, but also to perform the proposed Collaborative Contrastive Learning (C2L) strategy for learning high-level semantic consistency. Moreover, the C2LGM employs two respective decoders to reconstruct the complete image, and they leverage the structural spatial attention mechanism to facilitate the reconstruction of high-frequency structures.

Formally, given a corrupted image $I$, the encoder of C2LGM first encodes it into latent features of known content by iteratively stacked encoder blocks. Specifically, the detailed structure of encoder is composed of the residual blocks [47] to avoid overfitting and capture multi-level features from the input image:

$$F_{i+1} = f(F_i) + F_i,$$

where $F_i$ is the input feature of the $i$-th encoder block, and $f$ is the convolution and ReLU layers. To downsample the resolution of feature maps, we leverage the stride-2 convolution operation among resblocks.

1) STRUCTURAL KNOWLEDGE-GUIDED RECONSTRUCTION

Differing from methods [8] that employ a two-stage framework that first predicts the edge map, and then further leverages it to facilitate reconstructing the intact image, our C2LGM predicts the structural edge map and intact image, respectively. To be specific, the decoder $D^p$ is stacked by the structural spatial attention (SSA) modules, as shown in Figure 2. The SSA module predicts the structural attention map that contributes to inferring correct high-frequency
TABLE 1. Detailed architecture of C2LGM.

| Type     | Block       | Kernel | Stride | Outputs |
|----------|-------------|--------|--------|---------|
| Encoder  | Residual Block | 3x3    | 1      | 64      |
|          | Residual Block | 3x3    | 2      | 128     |
|          | Residual Block | 3x3    | 2      | 256     |
|          | Residual Block | 3x3    | 1      | 256     |
|          | SSA Module   | 3x3    | 1      | 256     |
|          | Upsample (×2↑) | -      | -      | -       |
|          | SSA Module   | 3x3    | 1      | 128     |
|          | Upsample (×2↑) | -      | -      | -       |
|          | SSA Module   | 3x3    | 1      | 64      |
|          | Convolution  | 3x3    | 1      | 3       |
| Decoder  | Channel Attention Block | 3x3 | 1 | 256 |
| D₁       | Upsample (×2↑) | -      | -      | -       |
|          | Channel Attention Block | 3x3 | 1 | 128 |
|          | Upsample (×2↑) | -      | -      | -       |
|          | Channel Attention Block | 3x3 | 1 | 64 |
|          | Convolution  | 3x3    | 1      | 1       |

structures of corrupted regions from feature maps in $D^p$:

$$A = f_{\text{att}}(\mathcal{H}(F^e)), \quad (2)$$

where $\mathcal{H}$ is the Channel-wise Attention Block (CAB) [48], $f_{\text{att}}$ is the spatial attention mechanism, and $A$ is the obtained structural attention map. Then, the high-frequency attention map $A$ is further leveraged as structural guidance for the content reconstruction in decoder $D^o$ by

$$F_j = \frac{f_{\text{att}}(\mathcal{H}(F_{j-1})) + \gamma A}{1 + \gamma} \otimes \mathcal{H}(F_{j-1}) + F_{j-1}, \quad (3)$$

where $\otimes$ is the Hadamard product, $F_j$ is the features of the $j$-th SSA module in decoder $D^o$, and $\gamma$ is a hyper-parameter to balance between two terms. In this way, the reconstructed content in image decoder $D^p$ is able to be guided by the edge decoder $D^e$, and such guidance ensures to inpaint reasonable high-frequency semantics.

B. COLLABORATIVE CONTRASTIVE LEARNING

Contrastive learning establishes positive and negative samples, and maps them into the same latent representations for feature learning. Differing from typical generative models that only perform pixel-level supervision on inpainted images, contrastive learning-based methods are able to cluster similar content and keep dissimilar content away in the latent space. However, the typical contrastive learning strategy samples image patches, and thus often fails to reconstruct reasonable semantic structures. Besides, the main point of contrastive learning is to establish appropriate contrastive pairs (positive or negative sample pairs). A mount of researches have proven the effectiveness of contrastive learning in recent high-level vision tasks [49], [50], [51], and obtain promising performance compared to typical fully supervised learning.

This is mainly because label-guided feature learning might decrease the quality of reconstruction for image inpainting. Besides, for low-level vision tasks, contrastive learning has not yet reached remarkable performance, which results from the difficulty of effectively learning reasonable semantic relations from established contrastive pairs. To this end, we propose a Collaborative Contrastive Learning (C2L) strategy to explore the potential of contrastive learning for the image inpainting task.

As illustrated in Figure 3, $C2LGM$ reuses its encoder to map the high-frequency (edge) patch and image patch into the same latent space and performs contrastive learning for synthesizing reasonable semantics in the corrupted regions. The main procedure of C2L is shown in Figure 3. Firstly, we crop the inpainted image, inpainted edge, groundtruth image, groundtruth edge, and randomly selected image patches and their aligned edge patches. Then we label the patch pairs from inpainted patches and groundtruth patches as positive samples, and label patch pairs of randomly selected images as negative pairs. Finally, the patches are fed into the encoder again for performing our collaborative contrastive learning strategy that learns the semantic consistency.

To be specific, we feed the query patches $x \in \mathbb{R}^{3 \times H \times W}$, the positive patches (image patch and aligned edge patch) $x^+ \in \mathbb{R}^{3 \times H \times W}$, and the negative patches (unrelated patches) $x^- \in \mathbb{R}^{3 \times H \times W}$ into the encoder to obtain feature representations, and then leverage two linear layers to map features into the same latent space and obtain K-dimension vectorial embeddings: query embeddings $z \in \mathbb{R}^{N_q \times K}$, positive embeddings $z^+ \in \mathbb{R}^{N_p \times K}$, and negative embeddings $z^- \in \mathbb{R}^{N_n \times K}$. Next, it performs noise contrastive estimation [52] (NCE) to cluster similar content by minimizing the distance of positive pairs, and keep dissimilar content away by maximizing the distance of negative pairs.

$$L_c = -\sum_{q=1}^{N_q} \log \left( \frac{\exp(z_q \cdot z_{q}^+)}{\sum_{p=1}^{N_p} \exp(z_q \cdot z_{p}^+) + \sum_{n=1}^{N_n} \exp(z_q \cdot z_{n}^-)} \right), \quad (4)$$

where $N_q$, $N_p$, and $N_n$ denote the number of query samples, positive samples, and negative samples, respectively. By performing the contrastive loss $L_c$, the positive samples are clustered away from the negative samples. Thus, $C2LGM$ is able...
Y. Du et al.: Collaborative Contrastive Learning-Based Generative Model for Image Inpainting

FIGURE 3. Illustration of the proposed collaborative contrastive learning approach. The positive pairs are established by patches from groundtruth, inpainted image, and their edge maps. The negative pairs are established by patches from randomly selected images and inpainted images. Besides, we perform contrastive loss by the latent vectors obtained by extra linear mapping \( L \) from features from the encoder \( E \). Note that \( z_q \), \( z_p \), and \( z_n \) are the query embedding, positive embedding, and negative embedding, respectively.

to synthesize reasonable semantic structures in accordance with known content.

C. TEXTURE DISTRIBUTION SAMPLING

Previous researches [26], [27], [53] suggest that there are many aspects of an image that can be regarded as stochastic, which will not affect human perception of the image when following the same distribution. Considering the difficulty of realizing such assumption, they add per-pixel noise after each convolution in generative adversarial networks to simulate such condition. However, the former analysis is insufficient since they randomly sample noise without strong prior knowledge. It’s hard to let random noise fit specific details of any image. Thus, we employ the texture distribution sampling approach to cope with such issues and improve the quality of image textures. The texture distribution sampling approach incrementally samples according to latent features from both known regions \( F_e \) (extracted by encoder) and inferred regions \( F_d \) (inferred by decoder). With the precise prediction of distribution for random content in the input image, our learned conditional prior leads to generating more reasonable content. Besides, such texture distribution sampling reduces fluctuation and content blurring, and also vivifies the image textures. In detail, the main goal of distribution sampling is to predict the conditional distribution of textures \( p(\text{noise}|x) \) in current prior knowledge \( x \). In each block, we predict the expectation and standard variation of normal distribution for conditional noise from current information. Inspired by recent improvements in hierarchy VAEs [28], [54], we combine features in the same scale from both encoder and decoder to predict texture distribution. In this way, our \( C2LM \) is able to model effective prior information for specific input images when sampling stochastic noise \( z \):

\[
\mu, \sigma = \text{Linear}(\alpha_e, \alpha_d),
\]

\[
z = \text{Sampling}(\mathcal{N}(\mu, \sigma^2)),
\]

where \( \mu \) and \( \sigma \) are the predicted mean and standard variation values, \( \mathcal{N}(\mu, \sigma^2) \) denotes normal distribution and Sampling is the random sampling operation from a known distribution. Following the way of training VAE, our \( C2LM \) leverages KL divergence between the conditional distribution and standard normal distribution to learn specific texture distribution for input images:

\[
\mathcal{L}_{\text{kl}} = w \cdot \text{KL}(\mathcal{N}(\mu, \sigma^2) | \mathcal{N}(0, 1)),
\]

where \( w \) is the empirical weight, and KL denotes the KL divergence that measures the distance between two distributions.

D. JOINTLY SUPERVISED PARAMETER LEARNING

We optimize the parameters of the proposed \( C2LM \) in an end-to-end manner, and perform supervision by three
main parts: the content reconstruction loss $L_{\text{rec}}$, high-level contrastive loss $L_{\text{c}}$, and the distribution loss of texture sampling $L_{\text{kl}}$.

The reconstruction loss includes the edge reconstruction loss $L_{\text{e}}$, content reconstruction loss $L_{\text{con}}$, and perceptual loss $L_{\text{p}}$:

$$L_{\text{rec}} = L_{\text{con}} + L_{\text{p}} + L_{\text{e}}. \quad (7)$$

1) **EDGE RECONSTRUCTION LOSS**

To synthesize accurate high-frequency content of corrupted regions, we perform edge reconstruction loss to minimize the $L_1$ distance between edges of output image $\hat{E}$ and corrupted image $E_{\text{gt}}$:

$$L_{\text{e}} = ||E_{\text{gt}} - \hat{E}||_1. \quad (8)$$

Specifically, the edge maps of groundtruth and output image are obtained by the canny algorithm [55] for edge detection.

2) **CONTENT RECONSTRUCTION LOSS**

Our C2LGM learns pixel-level reconstruction by employing $L_1$ distance between the reconstructed image and groundtruth image:

$$L_{\text{con}} = ||I_{\text{gt}} - \hat{I}||_1. \quad (9)$$

The perceptual loss [56] facilitates to learning semantic consistency between groundtruth image and reconstructed image, and it is defined as follows:

$$L_{\text{p}} = \sum_{l=1}^{L} ||f_{\text{vgg}}^l(\hat{I}) - f_{\text{vgg}}^l(I_{\text{gt}})||_1, \quad (10)$$

where $f_{\text{vgg}}^l(\hat{I})$ and $f_{\text{vgg}}^l(I_{\text{gt}})$ are the feature maps extracted by the $l$-th layer of pre-trained VGG-19 [57].

3) **GLOBAL-LOCAL ADVERSARIAL LOSS**

To make the reconstructed image as realistic as groundtruth images, we leverage two respective discriminators to perform the adversarial loss [44] on global image and local patches.

$$L_{\text{global}} = -E_{l=\text{global}}[D_\lambda(G(I))],$$
$$L_{\text{local}} = -E_{l=\text{local}}[D_\lambda(G(I))],$$
$$L_{\text{adv}} = L_{\text{global}} + L_{\text{local}}. \quad (11)$$

where $G$, $D_\lambda$, and $D_\phi$ are the proposed C2LGM and global-local discriminators, respectively. In this way, the discriminators minimize the distribution distance between reconstructed images and groundtruth images in both global and local views.

In sum, the complete loss functions of C2LGM are defined as follows:

$$\mathcal{L} = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{kl}} + \lambda_3 L_{\text{adv}}, \quad (12)$$

where $\lambda_1$, $\lambda_2$, and $\lambda_3$ are hyper-parameters to balance different losses. In our experiments, $\lambda_1 = 1$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.01$.

### IV. EXPERIMENTS

In this section, we conduct experiments to evaluate our proposed C2LGM both quantitatively and qualitatively. In Section IV-A, we describe the evaluation datasets, metrics, and implementation details. In Section IV-B, we conduct extensive experiments to compare our model with state-of-the-art methods for image inpainting. Finally, we perform ablation study to investigate the effect of each technique component in the proposed C2LGM in Section IV-C.

#### A. EXPERIMENTAL SETTING

1) **EVALUATION METRICS**

To measure the performance of our proposed C2LGM qualitatively and quantitatively in our experiments, we use four common evaluation metrics: the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index Measurement (SSIM) [58], the Local Mean Square Error (LMSE) [59], and the Learned Perceptual Image Patch Similarity (LPIPS) [60]. The PSNR is the ratio of maximum possible value to the power of distorting noise that affects the representation of quality between the reconstructed image and groundtruth:

$$\text{PSNR} = 20 \log_{10} \left( \frac{\text{MAX}_x^2}{\sqrt{\text{MSE}}} \right). \quad (13)$$

The SSIM models three factors, including luminance distortion, contrast distortion, and the loss of correlation, to measure the similarity between reconstructed image and groundtruth:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (14)$$

where $C_1$ and $C_2$ are constants. The LMSE calculates the squared intensity difference in a local perspective between the reconstructed image and groundtruth:

$$\text{LMSE} = \sum_{W_1, W_2} \text{MSE}(W_1, W_2), \quad (15)$$

where $W_1$ and $W_2$ are the local windows of reconstructed image and groundtruth. The LPIPS measures perceptual similarity by a pretrained network:

$$\text{LPIPS} = \sum \frac{1}{HW} \sum ||w \odot (\hat{y}_w - y_w)||. \quad (16)$$

where $\hat{y}_w$ and $y_w$ are local windows of reconstructed image and groundtruth, respectively.

Besides, we compare the visual results of inpainted images with state-of-the-art methods for image inpainting in the experiments. To avoid the bias of quantitative metrics, we also conduct human evaluation to investigate the performance of our C2LGM against other methods.

2) **DATASETS**

There are three benchmark datasets in image inpainting task used in our experiments:

- **Paris Street View** [61], which is captured from street views of Paris. By leveraging its official splits, we leverage 14,900 images for training and 100 images for testing.
TABLE 2. The experimental results of irregular image inpainting on the Paris Street View [61], CelebA-HQ256 [62], and Places [63] datasets. Higher PSNR and SSIM, lower LMSE and LPIPS denote better performance of image inpainting. The best results of each metric are highlighted in bold.

| Method          | Paris Street View | CelebA-HQ [62] | Places [63] |
|-----------------|-------------------|-----------------|-------------|
|                 | 20-30% 30-40% 40-50% | 20-30% 30-40% 40-50% | 20-30% 30-40% 40-50% |
| PC [5]          | 26.85 24.66 22.67 | 28.29 26.11 23.90 | 25.05 22.99 21.11 |
| GC [6]          | 26.29 23.99 21.86 | 27.81 25.52 23.19 | 25.15 23.00 20.94 |
| EdgeConnect [8] | 27.62 25.39 23.19 | 28.53 26.14 23.53 | 25.75 23.64 21.66 |
| LBAM [17]       | 27.54 25.25 23.09 | 28.86 26.55 24.22 | 25.69 23.46 21.41 |
| Jiang et al. [64] | 26.99 25.60 22.91 | 28.11 26.65 23.88 | 25.51 23.08 21.78 |
| LISK [10]       | 25.04 23.10 21.15 | 30.22 27.63 24.95 | 27.04 24.68 22.35 |
| E2I [65]        | 26.78 23.80 21.94 | 26.28 25.03 23.19 | 24.91 23.14 21.09 |
| GM-SRM [53]     | 29.04 26.79 24.51 | 30.23 27.91 25.51 | 27.09 24.93 22.91 |
| CTSDG [38]      | 27.33 24.67 22.00 | 28.49 25.99 23.77 | 26.35 24.11 21.58 |
| MEDPE [37]      | 27.23 24.78 22.30 | 25.50 24.10 22.31 | 24.84 23.38 21.38 |
| C2LGM (ours)    | 29.21 26.83 24.57 | 30.36 27.99 26.02 | 27.39 25.01 23.33 |

| Method          | PSNR | SSIM | LMSE | LPIPS |
|-----------------|------|------|------|-------|
| PC [5]          | 0.861| 0.795| 0.711| 0.844 |
| GC [6]          | 0.865| 0.801| 0.714| 0.842 |
| EdgeConnect [8] | 0.879| 0.821| 0.740| 0.805 |
| LBAM [17]       | 0.879| 0.816| 0.732| 0.807 |
| Jiang et al. [64] | 0.870| 0.812| 0.737| 0.809 |
| LISK [10]       | 0.881| 0.823| 0.744| 0.825 |
| E2I [65]        | 0.865| 0.810| 0.718| 0.845 |
| GM-SRM [53]     | 0.902| 0.852| 0.780| 0.825 |
| CTSDG [38]      | 0.874| 0.819| 0.728| 0.805 |
| MEDPE [37]      | 0.876| 0.811| 0.718| 0.884 |
| C2LGM (ours)    | 0.905| 0.853| 0.789| 0.826 |

- **CelebA-HQ** [62], which contains altogether 30,000 images of faces. By randomly selecting 3000 images as test dataset, we train our model by the remained 27000 images.
- **Places2** [63], which consists of over 2,000,000 images captured from 365 scene categories. In our experiments, We randomly selected 3,000 images as test set respectively, and the remained images are employed as training set.

Besides, for inpainting irregular corruption, we employ the mask dataset proposed by PC [5] to obtain irregularly corrupted images while training in different corrupted ratios.

3) IMPLEMENTATION DETAILS
In our experiments, we use 4 RTX 3080ti as computational devices for model training. The optimizer Adam [66] is used to optimize parameter learning during training with...
gradient descent optimization. For the hyper-parameters, the batchsize is set to be 4, learning rate is 0.0001, and overall epochs for training $C2LGM$ are 200. To augment training data, we resize images and randomly crop them into $256 \times 256$.

B. COMPARISON WITH STATE-OF-THE-ART METHODS

In this section, we compare our $C2LGM$ with current state-of-the-art methods for image inpainting on three benchmark datasets [61], [62], [63]. Particularly, we consider the change of corruption ratio in degraded images.

1) BASELINE METHODS

The competing baseline methods in our experiments include: (1) PC [5], which proposes the Partial Convolution (PC) for inpainting irregular corruption by learning to update the input mask in deep models; (2) GC [6], which learns the soft mask, denoted as the Gated Convolution (GC), like spatial attention mechanism to calculate the confidence of inpainted pixels; (3) EdgeConnect [8], which is a two-stage framework for image inpainting. It first predicts the edge map of the complete image, and then uses it to facilitate inpainting; (4) Jiang et al. [64], which leverages two respective discriminators to improve the performance of image inpainting; (5) LISK [10], which incorporates structural knowledge to reconstruct corrupted image and structure maps simultaneously; (6) E2I [65], which designs a two-step framework to first generate edges inside the missing areas, and then generate inpainted image based on the edges; (7) LBAM [17], which introduces a learnable reverse attention method for image inpainting; (8) CTSDG [38], which introduces a two-stream network to inpaint corrupted image by structural synthesis and texture reconstruction; (9) GM-SRM [53], which pretrains a generative memory to learn the semantic distribution of the whole dataset; (10) MEDFE [37], which introduces an encoder-decoder framework that learns to inpainting texture and structure information.

TABLE 3. The experimental results of regular image inpainting on the Paris Street View [61], CelebA-HQ256 [62], and Places [63] datasets. Higher PSNR and SSIM, lower LMSE and LPIPS denote better performance of image inpainting. The best results of each metric are highlighted in bold.

| Method         | Paris Street View | CelebA-HQ | Places2 |
|----------------|-------------------|-----------|---------|
|                | 25% | 50% | 25% | 50% | 25% | 50% |
| PC [5]         | 23.97 | 20.02 | 20.54 | 14.55 | 21.69 | 18.09 |
| GC [6]         | 23.32 | 19.88 | 25.45 | 19.86 | 21.27 | 17.54 |
| EdgeConnect [8] | 24.91 | 20.32 | 23.68 | 18.83 | 22.39 | 18.48 |
| LBAM [17]      | 24.18 | 20.14 | 25.91 | 20.73 | 21.79 | 18.38 |
| Jiang et al. [64] | 23.99 | 19.35 | 24.70 | 19.03 | 23.12 | 18.63 |
| LISK [10]      | 22.28 | 18.32 | 24.28 | 18.02 | 22.52 | 18.54 |
| E2I [65]       | 23.40 | 19.27 | 25.59 | 19.99 | 21.32 | 18.17 |
| GM-SRM [53]    | 25.86 | 21.41 | 26.97 | 21.75 | 23.33 | 19.51 |
| CTSDG [38]     | 24.22 | 19.67 | 25.51 | 20.28 | 22.87 | 18.40 |
| MEDFE [37]     | 23.43 | 19.09 | 25.54 | 20.26 | 21.18 | 17.57 |
| $C2LGM$ (ours) | 25.91 | 21.48 | 26.89 | 21.80 | 23.54 | 19.67 |

TABLE 4. The experimental results of irregular image inpainting by various methods on three benchmark datasets. (i.e., (20%, 30%], (30%, 40%], and (40%, 50%]), corruption types (regular or irregular), and four evaluation metrics (PSNR, SSIM, LMSE, and LPIPS). For a fair comparison, we use the same mask for each test image to evaluate different methods for image inpainting.

VOLUME 10, 2022
Besides irregular image inpainting, we also adopt center corruption to evaluate $C2LM$’s performance in Table 3 on three benchmark datasets in terms of four evaluation metrics. The results follow a similar effect as irregular inpainting, which again proves the effectiveness of $C2LM$ for image inpainting even on large continuous corruption. As shown in both Table 2 and Table 3, especially for the large corruption ratio, our $C2LM$ has a prominent improvement over other competing methods in four quantitative metrics, including PSNR, SSIM, LMSE, and LPIPS, which demonstrates the superiority of our method of coping with images with large corruption.

3) QUALITATIVE EVALUATION
In figure 4, we compare the results of our $C2LM$ with various state-of-the-art methods for image inpainting on randomly selected test samples from three benchmark datasets, including Paris Street View [61], CelebA-HQ [62], and Places2 [63]. Our $C2LM$ synthesizes the most realistic results than competing methods. Best viewed in zoom-in mode.

FIGURE 4. Visualization of results by our $C2LM$ and five state-of-the-art methods for image inpainting on seven randomly selected samples from the test set of Paris Street View [61], CelebA-HQ [62], and Places2 [63]. Our $C2LM$ synthesizes the most realistic results than competing methods. Best viewed in zoom-in mode.
blurriness, such as the first row in Figure 4(e). The way of learning both texture and structure by MEDFE [37] facilitates the reconstruction of corrupted content, but it often causes obvious semantic errors in Figure 4(f). On the contrary, the proposed C2LGM leverages contrastive learning to infer the semantics of corrupted content, and thus is able to synthesize more reasonable content than other state-of-the-art methods. Specifically, in the first and fourth rows in Figure 4, our C2LGM is able to reconstruct higher quality content that enjoys the most realistic details and reasonable semantics than other state-of-the-art methods for image inpainting from images with large corrupted areas. In summary, the experimental results demonstrate the excellent performance of our C2LGM for image inpainting, and it especially brings about fewer artifacts or blurriness in the reconstructed content while coping with images with large corrupted areas. This demonstrates again that C2LGM is able to synthesize higher quality content under the prior knowledge of structural information while reconstructing intact images.

C. ABLATION STUDY

To investigate each technique component in the proposed C2LGM, we perform ablation study in our experiments. In detail, there are altogether four variants of the proposed C2LGM:

- **EDM**, which employs a plain Encoder-Decoder based Model to inpaint corrupted images. No contrastive learning or edge learning is adopted;
- **SKR**, which further leverages Structural Knowledge-guided Reconstruction as high-frequency guidance for inpainting content of corrupted regions;
- **TDS**, which employs the Texture Distribution Sampling approach to improve the quality of image details;
- **CLGM**, which is a Contrastive Learning Generative Model that leverages plain contrastive learning mechanism to keep semantic consistency between inferred content and known content;
- **C2LGM**, which further leverages the proposed Collaborative Contrastive Learning Generative Model to synthesize the intact image from the corrupted image. The resulting model is the complete model C2LGM.

Table 4 lists the results of four variants of C2LGM on CelebA-HQ dataset [62] in terms of four evaluation metrics, considering two corruption ratios.

1) EFFECT OF STRUCTURAL KNOWLEDGE-GUIDED RECONSTRUCTION

In Table 4, a prominent gap between EDM and SKR proves the necessity of a structural knowledge-guided reconstruction branch. There is a significant improvement in inpainting performance compared to a plain encoder-decoder based framework. This is reasonable that more accurate high-frequency information leads to a higher quality of content synthesis.

2) EFFECT OF TEXTURE DISTRIBUTION SAMPLING

Though the quantitative improvement between SKR and TDS is not too large, the improvement in visual quality can be found in Figure 5. By sampling stochastic content by predicted distribution, the detail quality of synthesized content can be further enhanced.

3) EFFECT OF CONTRASTIVE LEARNING

Comparing the performance between TDS and CLGM, it manifests that contrastive learning facilitates learning semantic consistency between synthesized content and known content.
known content. Contrastive learning pushes the similar semantics close and dissimilar content far apart. Such mechanism is able to improve the performance of image inpainting.

4) EFFECT OF COLLABORATIVE CONTRASTIVE LEARNING
The comparison between CLGM and C2LGM in Table 4 demonstrates that our proposed collaborative contrastive learning effectively combines high-frequency information with pixel-level information. Such way of establishing contrastive pairs enables deep models to enjoy more accurate consistency of inferred and known semantics.

5) QUALITATIVE ABLATION EVALUATION
As shown in Figure 5, inpainted results of five variants of C2LGM are visualized. The results show increasingly better quality of inpainted content as the augment of C2LGM with incremental functional components.

6) HUMAN EVALUATION
The comparison by metrics has their bias for accurate evaluation of synthesized content. Thus, we perform human evaluation on three competing state-of-the-art methods for image inpainting to avoid bias, including PC [5], EdgeConnect [8], and MEDFE [37]. There are 50 images that are randomly selected from the test datasets, and their results are sent to 50 human judges for ranking the reasonability of different methods. The voting results on two benchmark datasets are listed in Table 5, and our C2LGM obtains the best votes that are prominently higher than competing state-of-the-art methods. It demonstrates again our proposed C2LGM improves the visual quality of synthesized content, even coping with large corruptions.

7) FAILURE CASES
Even if our C2LGM has shown significant performance while reconstructing images with large continuous corruption in image inpainting, negative results are observed for a few challenging cases as well. As shown in Figure 6, our C2LGM fails to synthesize plausible content when the objects are largely occluded. Although our C2LGM attempts to synthesize more realistic details, too limited information leads to inferring incomplete semantics in the corrupted content, such as the paddles in the first image and the man on the right in the second image. Such challenging conditions result in performance degradation and in our future work, we plan to solve these problems.

V. CONCLUSION
In this paper, we propose a collaborative contrastive learning-based framework C2LGM that learns the semantic consistency between the inferred content and known content for image inpainting. The proposed collaborative contrastive learning establishes the positive pairs and negative pairs with the assistance of high-frequency edge maps, and thus deep models are able to synthesize reasonable semantic content. Besides, by texture distribution sampling, the inferred content enjoys realistic details from the distribution of known content. Consequently, our C2LGM is able to infer higher quality content for corrupted regions, particularly large continuous corruption.

REFERENCES
[1] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, “GLEAN: Generative latent bank for large-factor image super-resolution,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 14245–14254.
[2] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan, “Deep joint rain detection and removal from a single image,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1357–1366.
[3] S. J. J. Ren, L. Xu, Q. Yan, and W. Sun, “Shepard convolutional neural networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 1–9.
[4] D. Pathak, P. Krishnan, 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.
[5] 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., 2018, pp. 85–100.
[6] 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.
[7] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, “Recurrent feature reasoning for image inpainting,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 7760–7768.
[8] K. Nazari, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, “EdgeConnect: Structure guided image inpainting using edge prediction,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW). IEEE, 2019, pp. 3265–3274.
[9] 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.
[10] J. Yang, Z. Qi, and Y. Shi, “Learning to incorporate structure knowledge for image inpainting,” in Proc. Assoc. Advan. Artif. Intell. (AAAI), 2020, vol. 34, no. 7, pp. 12605–12612.
[11] 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. 5962–5971.
[12] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C.-C. J. Kuo, “Contextual-based image inpainting: Infer, match, and translate,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.
[13] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, “Shift-Net: Image inpainting via deep feature rearrangement,” in Proc. Eur. Conf. Comput. Vis., Sep. 2018, pp. 1–17.
[14] 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.
[15] N. Wang, J. Li, L. Zhang, and B. Du, “MUSICAL: Multi-scale image contextual attention learning for inpainting,” in Proc. 28th Int. Joint Conf. Artif. Intell., Aug. 2019, pp. 3748–3754.

[16] 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.

[17] C. Xie, S. Liu, C. Li, M.-M. Cheng, W. Zuo, X. Liu, S. Wen, and E. Ding, “Image inpainting with learnable bidirectional attention maps,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8858–8867.

[18] N. Wang, S. Ma, J. Li, Y. Zhang, and L. Zhang, “Multistage attention network for image inpainting,” Pattern Recognit., vol. 106, Oct. 2020, Art. no. 107448.

[19] 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.

[20] A. A. Efros and W. T. Freeman, “Image quilting for texture synthesis and transfer,” in Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn., 2001, pp. 341–346.

[21] V. Kwatra, I. Essa, A. Bobick, and N. Kwatera, “Texture optimization for example-based synthesis,” ACM Trans. Graph., vol. 24, no. 3, pp. 795–800, 2005.

[22] A. Vasswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30. Red Hook, NY, USA: Curran Associates, 2017, pp. 1–11.

[23] 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.

[24] L. Zhao, Q. Mo, S. Lin, Z. Wang, Z. Zuo, H. Chen, W. Xing, and D. Lu, “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.

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

[26] 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. 4170–4179.

[27] T. Karras, S. Laine, and T. Aila, “Progressive growing of GANs,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 5880–5884.

[28] H. Roy, S. Chaudhury, T. Yamasaki, and T. Hashimoto, “Image inpainting via conditional texture translation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 795–802.

[29] A. Razavi, A. Van Den Oord, and O. Vinyals, “Generating diverse high-fidelity images with VQ-V AE-2,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–11.

[30] X. Feng, W. Pei, Z. Jia, F. Chen, D. Zhang, and G. Lu, “Deep-masking generative network: A unified framework for background restoration from superimposed images,” 2020, arXiv:2010.04324.

[31] D. Bo, R. Ma, K. Wang, M. Su, and P. An, “Deep learning-based image inpainting with structure map,” J. Electron. Imag., vol. 30, no. 2, Apr. 2021, Art. no. 023016.

[32] H. Ji, X. Feng, W. Pei, J. Li, and G. Lu, “U2-FORMER: A nested U-shaped transformer for image restoration,” 2021, arXiv:2112.02279.

[33] I. J. Goodfellow, P. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proc. 27th Int. Conf. Neural Inf. Process. Syst., vol. 2, 2014, pp. 2672–2680.

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

[35] A. Romero, A. Castillo, J. Abril-Nova, R. Timofte, R. Das, S. Hira, Z. Pan, M. Zhang, B. Li, D. He, and T. Lin, “NTIRE 2022 image inpainting challenge: Report,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), Jun. 2022, pp. 1150–1182.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[37] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[38] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9729–9738.

[39] O. Henaff, “Data-efficient image recognition with contrastive predictive coding,” in Proc. ICML, 2020, pp. 4182–4192.

[40] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in Proc. ICLR, 2020, pp. 1597–1607.

[41] A. Van Den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2018, arXiv:1807.03748.

[42] X. Feng, W. Pei, L. Li, F. Chen, D. Zhang, and G. Lu, “Generative memory-guided semantic reasoning model for image inpainting,” 2021, arXiv:2110.00261.

[43] A. Razavi, A. Van Den Oord, and O. Vinyals, “Generating diverse high-fidelity images with VQ-VAE-2,” in Proc. Adv. Neural Inf. Process. Syst., vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 1–11.

[44] J. Canny, “A computational approach to edge detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[45] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2016, pp. 694–711.

[46] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, arXiv:1409.1556.

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

[48] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, “Ground truth dataset and baseline evaluations for intrinsic image algorithms,” in Proc. IEEE 12th Int. Conf. Comput. Vis., Sep. 2009, pp. 2335–2342.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 585–595.

[50] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, “What makes Paris look like Paris?” Commun. ACM, vol. 58, no. 12, pp. 103–110, Nov. 2015.

[51] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in Proc. Int. Conf. Learn. Represent., 2018, pp. 1–26.
Y. Du et al.: Collaborative Contrastive Learning-Based Generative Model for Image Inpainting

[63] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, “Places: A 10 million image database for scene recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1452–1464, Jun. 2018.

[64] Y. Jiang, J. Xu, B. Yang, J. Xu, and J. Zhu, “Image inpainting based on generative adversarial networks,” IEEE Access, vol. 8, pp. 22884–22892, 2020.

[65] 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.

[66] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, arXiv:1412.6980.

YONGQIANG DU received the bachelor’s degree in agricultural machinery design and manufacturing from the Luoyang Institute of Technology, in 1992, and the master’s degree in computer application from Jiangsu University, Zhenjiang, China, in 2005. Since 2006, he has been an Associate Professor with Xinyang Agriculture and Forestry University. His research interests include computer graphics, pattern recognition, and intelligent systems.

HAORAN LIU received the bachelor’s degree in engineering mechanics from the Changsha University of Science and Technology, in 2016, and the M.Sc. degree in mechanical and aerospace engineering from George Washington University, Washington, DC, USA, in 2019. Since 2020, he has been a Teaching Assistant with Xinyang Agriculture and Forestry University. His research interests include computer graphics and intelligent controls.

SONGNAN CHEN received the B.S. degree in computer science and technology from Xinyang Normal University, Henan, in 2010, the M.S. degree in computer technology from Sun Yat-Sen University, Guangzhou, in 2014, and the Ph.D. degree in forest engineering from Beijing Forestry University, Beijing, China, in 2021. Currently, he is a Lecturer with the School of Mathematics and Computer Science, Wuhan Polytechnic University, China. His research interests include digital image processing, machine learning, computer vision, and deep learning.

***