PGI: A prior-guided inversion model for image inpainting with enhanced textures and less artifacts

1st Zixuan Chen
School of Cyber Security and Computer
Hebei University
Baoding, Hebei Province, China
chenzixuan@stumail.hbu.edu.cn

2nd Liang Wang*
School of Cyber Security and Computer
Hebei University
Baoding, Hebei Province, China
wangl@hbu.edu.cn

3rd Shaokang Zhang
School of Cyber Security and Computer
Hebei University
Baoding, Hebei Province, China
zhangshaokang@hbu.edu.cn

***Abstract*—Deep image inpainting has made significant strides due to recent advancements in image generation and processing algorithms. However, effectively handling fine-grained textures and structural details while avoiding noticeable artifacts remains an unsolved problem. The challenge lies in the uncertainty within the image generation process, where mapping degraded inputs to images inevitably leads to degradation without effective prior guidance. In view of this, we propose a prior-guided inversion model (PGI) that systematically integrates both learning-based and optimization-based inversion strategies. Specifically, generative priors are first obtained and refined through a learning-based inversion, and then enhanced their textures with our designed GFFM and W-mix module. The enhanced priors are then further utilized to guide an optimization-based inversion for improved texture generation. Experimental results on widely used public datasets show that PGI effectively alleviates information loss and produces more natural and coherent restored images compared with existing representative methods.**

***Index Terms*—Prior-guided inversion; latent codes mixing; multi-stage expansion generative; StyleGAN invension; image inpainting**

I. Introduction

Image inpainting is a fundamental computer vision task, aiming to recover missing or damaged regions while preserving texture, color, and structural consistency. It has broad applications, such as reconstructing missing areas in MRI or X-rays for medical diagnosis [1], and restoring incomplete satellite images caused by clouds or sensor failures [2].

Existing interpolation- and texture synthesis–based methods are fundamentally constrained in their ability to generalize to unseen textures [3]. While deep learning–driven approaches—particularly diffusion-based [4] and GAN-based [5] methods—have markedly advanced image generation capabilities, they remain hampered by persistent limitations. Specifically, diffusion-based methods often entail computationally expensive iterative sampling and may produce outputs with structural inconsistencies; GAN-based methods, in turn, frequently underperform in reconstructing fine-grained geometric structures and high-frequency textural details. These shortcomings largely arise from the intrinsic ambiguity of inverse mapping: recovering photo-realistic images from degraded inputs is inherently ill-posed without strong, well-structured prior constraints.

To achieve image inpainting with enhanced textures and less artifacts, we propose a prior-guided inversion model (PGI) based on StyleGAN2 inversion. The main contributions of this work are summarized as follows:

1) A Fast Fourier Style 2 (FFS2) encoder is built to adaptively prepare features for subsequent processing by Fast Fourier Convolution (FFC) [6], enabling more effective extraction and utilization of multi-level representations. It improves semantic consistency and enhances texture representation during the inversion process.
2) The mixing latent codes obtained from both learning- and optimization-based inversion strategies are verified to be capable of effectively alleviating the information loss and irreversibility issues inherent in single inversion schemes. They enable more accurate and stable image reconstruction.
3) A two-stage expansion generation process guided by the mixed latent representation is proposed to reduce edge artifacts and texture distortion commonly observed in inversion-based methods. It yields more natural and structurally coherent inpainting results.

II. Related work

A. Generative adversarial networks (GAN)

GAN-based models for image inpainting have been widely adopted in recent years. Co-Mod-GAN [7] collaboratively modulates both the conditional input and random latent vector at each convolutional layer, improving model training across multiple modalities, and enabling it to understand and generate various forms of input data. This not only enhances the quality of generated images but also increases adaptability to unseen modalities. CM-GAN [8] incorporates Fourier convolution blocks and a decoder with global spatial modulation blocks into GAN, using an encoder to extract multi-scale feature representations from input images with holes, thereby facilitating better synthesis of both overall structure and local details. However, these methods do not address the issue of the high computational resources required for GAN training. We aim to alleviate this issue by focusing on editing and

*Corresponding author. This research was supported by the Fund of Central Government for Local Science and Technology Development (246Z0704G).

improving pretrained models. Therefore, due to the excellent performance of StyleGAN in image processing, we follow established methods and select StyleGAN2 as inversion target in this paper.

B. GAN inversion

GAN inversion aims to map a given image back into the latent space of a pretrained GAN model. The generator can then faithfully reconstruct the image from the inverted code. GAN inversion can be categorized into two types of methods: learning- and optimization-based GAN inversions. Learning-based methods train encoders to predict latent codes from input images. Optimization-based methods iteratively refine the latent codes through backpropagation to minimize pixel-wise reconstruction loss. The hybrid of the two methods is also appeared in some studies. However, the analysis of existing research shows that GAN inversion performs poorly in preserving the original image's semantic content and fine-grained textures [9], which is the focus of our work.

III. Method

For a damaged image, PGI works in two stages, as shown in Fig. 1. In the first stage, we design an FFS2 encoder to extract and stabilize features at multiple scales, and employ a latent code search strategy that combines the learning capability of the FFS2 encoder with the generation optimization of StyleGAN2. This strategy enables more effective extraction of informative representations from real images. In the second stage, an optimization-based image inversion strategy is adopted. Specifically, the latent codes obtained in the first stage are fused with latent representations derived from average-vector–based expansion optimization through a Global Fast Fourier Convolution Module (GFFM) and a W-mix module. The fused latent representation is then fed into the generator for further expansion optimization, leading to the reconstruction of complete images.

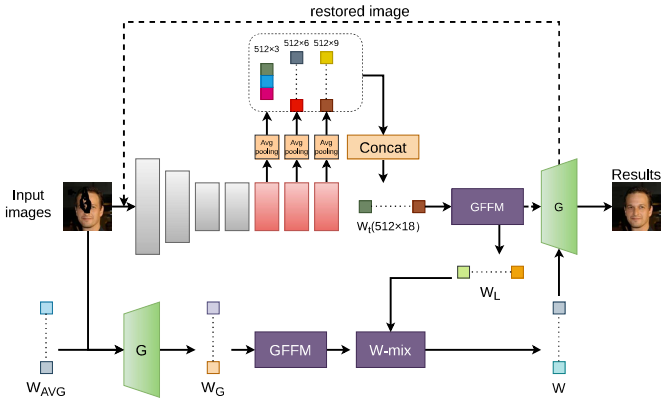

Fig. 1. PGI working flow.

A key design of PGI is the prior-guided inversion mechanism. Specifically, PGI treats the structured latent space learned by the encoder as a generative prior to constrain subsequent latent codes searching and optimization-based image generation, thereby effectively reducing the solution space of the inversion problem. This prior constraint helps preserve semantic consistency during image reconstruction, facilitating reasonable structural recovery and coherent texture synthesis.

A. FFS2 encoder

Table I summarizes the structure of the FFS2 encoder, which consists of seven convolutional modules. The first four modules progressively increase the number of channels to perform hierarchical feature abstraction, providing a stable low- to mid-level representation of the input image. Based on this backbone, features are extracted from the last three modules for style code prediction. In this way, we divide features from different depths into three groups, enabling the encoder to model global structural information, mid-level semantic attributes, and fine-grained texture details at different network stages. Moreover, to ensure robustness to spatial variations and alleviate instability caused by feature scale differences, average pooling is applied during style vector regression. Finally, the style vectors predicted at each selected module are concatenated to form a complete StyleGAN latent codes set. This hierarchical grouping and pooling stabilize the style regression and ensure the encoder captures complementary information across multiple scales.

TABLE I
FFS2 Encoder Structure

Module name	Kernel	Filters	BatchNorm	Non-linearity
Conv1	7×7	64	–	ReLU
Conv2	5×5	128	✓	ReLU
Conv3	5×5	256	✓	ReLU
Conv4	3×3	512	✓	ReLU
Conv5	3×3	512	✓	PReLU
Conv6	3×3	512	✓	PReLU
Conv7	3×3	512	✓	PReLU

B. Global Fast Fourier Convolution Module

To better exploit the relationships among features from different modules, we introduce FFC into the inpainting model, motivated by its strong capability in capturing global structural information. As illustrated in Fig. 2, the feature processing is referred as to GFFM, which operates as follows:

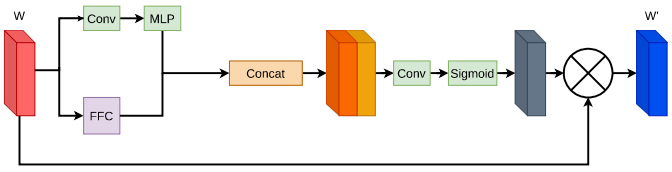

Fig. 2. GFFM module.

First, feature maps received from the convolutional backbone are processed along the channel dimension through parallel MLP and FFC branches, and the resulting representations are concatenated. Subsequently, a convolution operation with a kernel size of 5×5 is applied to the fused features, followed by a Sigmoid function to generate a global weighting map.

Finally, the output feature map is obtained by performing element-wise multiplication between the weighting map and the input features. The process can be formulated as follows:

$$F_{\text{fusion}} = \text{Concat}\left(\text{MLP}\left(\text{Conv}_{1\times1}\left(W\right)\right), \text{FFC}\left(W\right)\right), \quad (1)$$

$$F = \sigma\left(\text{Conv}_{5\times5}\left(F_{\text{fusion}}\right)\right), \quad (2)$$

$$W' = W \odot F. \quad (3)$$

By assigning higher priorities to informative regions in the image – such as high-frequency textures and critical structural components – the proposed module effectively enhances both the accuracy and efficiency of fine-detail image inpainting.

C. W-mix

In image inpainting, balancing features from different sources remains a fundamental challenge. Latent codes obtained through learning-based inversion are directly derived from the original image and therefore preserve more source-specific information, whereas optimization-based latent codes typically achieve higher generation quality but are not strictly constrained by the original image content. However, simply concatenating the two inversion branches often leads to unreasonable blurring artifacts due to the non-adaptability among heterogeneous features.

To effectively exploit multi-source latent representations for suppressing noise, blur, and other interference factors, we propose a feature fusion module, named W-mix, that enables stable and adaptive integration of heterogeneous latent codes, as illustrated in Fig. 3.

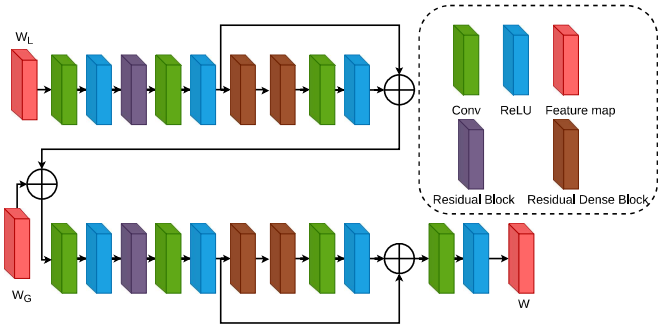

Fig. 3. W-mix feature fusion module.

The proposed design not only stabilizes fine-grained texture details but also maximizes the complementary strengths of different inversion strategies. Specifically, residual connections and dense residual links are employed to facilitate the stable propagation and fusion of local and global features, thereby ensuring texture consistency during the generation process. The process $f(\cdot)$ can be represented as:

$$f\left(\cdot\right) = \phi_{\text{ReLU}}\left(\text{Conv}_2\left(\text{RDB}\left(\text{RB}\left(\phi_{\text{ReLU}}\left(\text{Conv}_1\left(\cdot\right)\right)\right)\right)\right)\right). \quad (4)$$

Furthermore, to promote effective interaction between the two latent representations, the processed latent codes $\mathbf{W}_L$ is used to guide the refinement of $\mathbf{W}_G$, enabling more reliable regulation of latent quality and improving overall restoration fidelity. The process can be formulated as follows:

$$W_L' = f(W_L), \quad (5)$$

$$W_1 = W_L' \oplus W_G, \quad (6)$$

$$W = f(W_1). \quad (7)$$

The W-mix module adaptively fuses heterogeneous latent codes, effectively reducing blurring, noise, and other artifacts while preserving fine-grained textures and structural details. Its residual and dense connections ensure stable feature propagation, facilitating complementary information from both inversion branches.

D. Loss functions

Common Losses. To ensure accurate pixel-wise reconstruction between the predicted image and the ground truth, we employ the L_1, L_2, and LPIPS loss functions [23].

Multi-Scale Structure-Aware Edge Loss. We propose a structure-aware edge loss to enhance structural consistency by emphasizing strong gradient regions using a parameter-free weighting scheme:

$$L_{\text{MSE}} = \sum_{s\in S} \lambda_s \sum_{i,j} w_s(i,j) \left|\nabla x_s(i,j) - \nabla \hat{x}_s(i,j)\right|, \quad (8)$$

$$w_s(i,j) = \frac{\|\nabla x_s(i,j)\|_2}{\mathbb{E}\left[\|\nabla x_s\|_2\right] + \epsilon}, \quad (9)$$

where x_s and $\hat{x}_s$ denote the ground-truth and reconstructed images at scale s, and $\nabla(\cdot)$ computes image gradients.

The overall definition of our loss function is:

$$L_{\text{total}} = \lambda_1 \cdot L_1 + \lambda_2 \cdot L_2 + \lambda_3 \cdot L_{\text{LPIPS}} + \lambda_4 \cdot L_{\text{MSE}}, \quad (10)$$

where $\lambda_1 = 1, \lambda_2 = 0.8, \lambda_3 = 0.5, \lambda_4 = 0.03$ are hyper-parameters that balance the contributions of each loss component.

IV. Experiments

A. Datasets and Implementation Details

To validate the effectiveness and generalization ability of PGI, we conducted experiments on several publicly available image inpainting benchmark datasets, including CelebA [10], Places2 [11], and Paris Street View [12]. All datasets were uniformly preprocessed to a resolution of 256×256 pixels, and both training and testing were conducted at this resolution.

For masks, we adopted the mask dataset proposed by Liu et al. [13]. This mask dataset contains missing regions of various shapes and scales, with a resolution of 256×256 pixels.

During training and evaluation, the model was trained and tested separately on the CelebA, Places2, and Paris Street View datasets. The model was trained for 100 epochs. During the testing phase, to systematically evaluate the inpainting performance under different levels of missing regions, masks from five missing-area ratio intervals were applied to the test images, namely 10%–20%, 20%–30%, 30%–40%, 40%–50%, and 50%–60%.

B. Baselines and metrics

To qualitatively and quantitatively evaluate the inpainting performance of PGI, we compared PGI with several representative image inpainting methods, including EC [14], LaMa [15], RePaint [16], Robust Unsupervised(RU) [17], Spa-Former [18], and HINT [19]. For quantitative evaluation, several widely used metrics were adopted, including SSIM [20], PSNR [21], FID [22], and LPIPS [23]. All experiments were implemented using the PyTorch framework. The training process was performed on four NVIDIA A10 GPUs (30 GB), with the learning rate set to 0.0001. All results are obtained by averaging over multiple independent runs with different random seeds.

C. Quantitative Comparison under Different Occlusion Ratios

Table II to Table IV present quantitative comparisons of different methods on the CelebA, Places2, and Paris Street View datasets under varying mask ratios. The experimental results show that when the missing regions are relatively small, the proposed method achieves performance comparable to or surpassing mainstream approaches. As the missing area gradually increases, our method demonstrates increasingly clear performance advantages, highlighting its stronger sensitivity to structural and texture information as well as its ability to better preserve global semantics and information completeness.

Fig. 4 to Fig. 6 present qualitative comparisons of image inpainting results produced by different methods under varying mask ratios on the CelebA, Places2, and Paris Street View datasets. As can be observed, the compared methods exhibit noticeable differences in terms of structural reconstruction, texture continuity, and the naturalness of transitions between the restored and known regions. Overall, methods such as EC, Robust Unsupervised, and Spa-Former show limitations in certain scenarios, including incomplete structures, discontinuous textures, or visible artifacts. In contrast, the proposed method is able to generate more coherent and visually detailed inpainting results for most test cases, demonstrating advantages in texture preservation and artifact suppression. In some challenging cases, slight blurring can still be observed when restoring some fine-grained semantic structures, such as textual patterns.

TABLE II
QUANTITATIVE RESULTS ON THE CELEBA DATASET

Mask Ratio	Metric	EC	LaMa	RePaint	RU	Spa-Former	HINT	Ours
10%–20%	SSIM ↑	0.934	0.968	0.943	0.953	0.944	0.950	**0.972**
	PSNR ↑	31.97	32.25	30.93	32.42	33.68	33.29	**34.81**
	FID ↓	5.74	3.91	5.23	4.78	2.93	3.20	**2.91**
	LPIPS ↓	0.067	0.064	**0.048**	0.054	0.062	0.058	**0.048**
20%–30%	SSIM ↑	0.911	0.921	0.884	0.923	0.919	0.916	**0.929**
	PSNR ↑	28.68	29.28	28.66	27.69	30.45	30.60	**30.71**
	FID ↓	5.42	4.73	9.50	**4.12**	3.95	5.30	5.23
	LPIPS ↓	0.082	0.065	0.075	0.097	0.082	0.081	**0.063**
30%–40%	SSIM ↑	0.867	0.899	0.878	0.874	0.894	0.876	**0.901**
	PSNR ↑	28.14	26.85	27.63	24.67	27.91	27.94	**28.41**
	FID ↓	10.41	7.62	16.34	13.41	**5.35**	7.41	7.86
	LPIPS ↓	0.117	0.110	0.107	0.123	0.114	0.122	**0.104**
40%–50%	SSIM ↑	0.817	0.830	0.806	0.823	0.851	**0.853**	**0.853**
	PSNR ↑	24.10	25.00	24.61	22.87	26.00	25.74	**26.10**
	FID ↓	11.01	**6.38**	22.61	15.83	7.29	10.16	10.41
	LPIPS ↓	0.157	0.136	0.133	0.151	0.120	0.143	**0.118**
50%–60%	SSIM ↑	0.735	0.742	0.754	0.759	0.832	0.769	**0.835**
	PSNR ↑	21.51	**24.69**	23.71	21.95	24.19	23.81	24.32
	FID ↓	16.45	11.28	15.70	17.87	**7.52**	13.32	12.80
	LPIPS ↓	0.216	0.205	0.211	0.221	0.148	0.178	**0.145**

TABLE III
QUANTITATIVE RESULTS ON THE PLACES2 DATASET

Mask Ratio	Metric	EC	LaMa	RePaint	RU	Spa-Former	HINT	Ours
10%–20%	SSIM ↑	0.930	0.947	0.942	0.911	0.941	0.954	**0.958**
	PSNR ↑	28.85	28.28	29.06	28.95	29.19	29.88	**30.04**
	FID ↓	10.46	14.47	13.51	14.87	10.36	13.91	**8.90**
	LPIPS ↓	0.056	0.049	0.047	0.065	0.050	0.044	**0.039**
20%–30%	SSIM ↑	0.805	0.893	0.842	0.842	0.901	0.909	**0.915**
	PSNR ↑	21.25	24.77	20.49	25.91	25.83	**27.33**	27.09
	FID ↓	24.14	24.62	33.15	21.05	17.68	19.83	**16.65**
	LPIPS ↓	0.079	0.071	0.227	0.103	0.076	0.076	**0.071**
30%–40%	SSIM ↑	0.807	0.823	0.701	0.789	0.849	0.856	**0.868**
	PSNR ↑	22.42	22.37	20.71	24.04	23.47	25.37	**25.51**
	FID ↓	22.13	34.85	32.34	29.08	25.31	**20.02**	21.93
	LPIPS ↓	0.166	0.168	0.224	0.211	0.115	0.114	**0.111**
40%–50%	SSIM ↑	0.728	0.742	0.746	0.762	0.776	0.783	**0.812**
	PSNR ↑	21.18	20.58	23.68	20.60	21.65	23.71	**23.84**
	FID ↓	14.95	44.63	20.85	35.60	33.81	32.12	**19.76**
	LPIPS ↓	0.211	0.182	0.242	0.252	0.164	0.158	**0.153**
50%–60%	SSIM ↑	0.718	0.722	0.711	0.641	0.745	0.732	**0.762**
	PSNR ↑	20.28	19.20	19.07	20.42	21.10	21.46	**21.80**
	FID ↓	23.08	25.93	24.98	25.08	**22.41**	25.71	22.98
	LPIPS ↓	0.244	0.219	0.267	0.305	0.235	0.244	**0.216**

TABLE IV
QUANTITATIVE RESULTS ON THE PARIS STREET VIEW DATASET

Mask Ratio	Metric	EC	LaMa	RePaint	RU	Spa-Former	HINT	Ours
10%–20%	SSIM ↑	0.916	0.951	0.943	0.933	0.938	0.965	**0.971**
	PSNR ↑	30.37	32.49	29.41	31.08	31.75	32.08	**33.01**
	FID ↓	24.06	18.54	15.71	16.08	13.02	12.87	**11.98**
	LPIPS ↓	0.068	0.036	0.054	0.068	0.074	**0.033**	**0.033**
20%–30%	SSIM ↑	0.884	0.907	0.903	0.872	0.923	0.934	**0.936**
	PSNR ↑	26.04	27.41	27.11	25.87	28.64	29.14	**29.65**
	FID ↓	28.31	35.07	24.50	26.76	23.65	21.41	**21.32**
	LPIPS ↓	0.112	0.098	0.077	0.074	0.071	0.056	**0.054**
30%–40%	SSIM ↑	0.832	0.853	0.861	0.835	0.855	0.888	**0.893**
	PSNR ↑	25.16	25.46	26.02	25.81	27.35	26.37	**27.60**
	FID ↓	52.81	48.82	48.09	37.76	35.64	32.31	**31.86**
	LPIPS ↓	0.144	0.106	0.122	0.124	0.107	0.089	**0.085**
40%–50%	SSIM ↑	0.771	0.782	0.791	0.789	0.807	0.831	**0.851**
	PSNR ↑	21.24	23.65	24.09	22.98	25.45	24.74	**26.26**
	FID ↓	72.76	68.48	61.63	58.89	47.83	45.54	**41.20**
	LPIPS ↓	0.209	0.134	0.155	0.176	0.182	0.138	**0.123**
50%–60%	SSIM ↑	0.688	0.723	0.727	0.711	0.704	0.715	**0.743**
	PSNR ↑	19.91	20.17	21.87	21.17	22.19	22.31	**24.72**
	FID ↓	85.93	67.31	83.89	65.87	56.64	54.41	**52.87**
	LPIPS ↓	0.224	0.194	**0.167**	0.231	0.248	0.219	0.187

D. Ablation Study Analysis

1) Effectiveness of Modules: Table V evaluates different module combinations under varying mask ratios. The baseline performs poorly, with accuracy degrading as mask size increases. Adding GFFM yields consistent but limited gains,

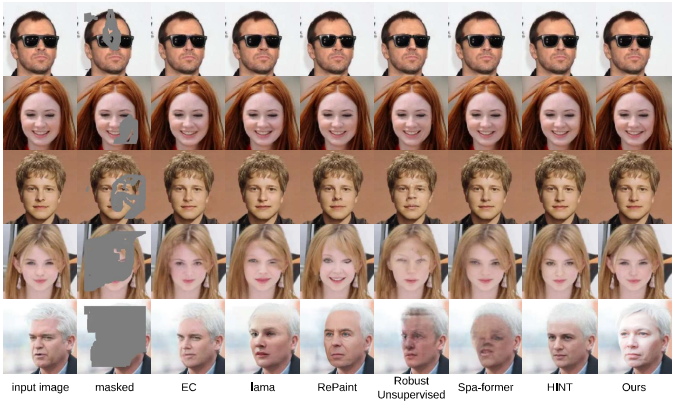

Fig. 4. Results on CelebA.

Fig. 5. Results on Place2.

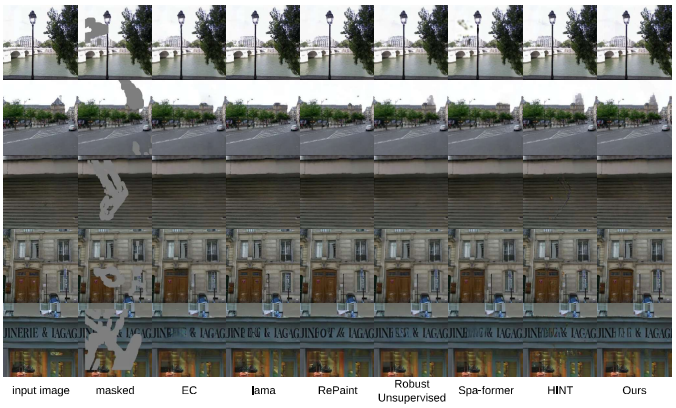

Fig. 6. Results on Paris Street View.

while W-mix brings more significant improvements, especially under medium and large masks. Combining both modules achieves the best performance across all settings, confirming their complementary effects on robustness and reconstruction quality. Among all components, W-mix contributes the most significant performance gain, validating its central role in bridging hybrid inversion.

TABLE V
ABLATION STUDY ON KEY MODULES

Modules	Metrics	10–20%	20–30%	30–40%	40–50%	50–60%
none	SSIM ↑	0.923	0.897	0.868	0.824	0.781
	PSNR ↑	28.62	26.91	25.14	23.02	21.37
	FID ↓	7.32	9.86	13.41	17.83	21.76
	LPIPS ↓	0.124	0.146	0.178	0.216	0.254
GFFM	SSIM ↑	0.948	0.913	0.884	0.846	0.808
	PSNR ↑	31.47	28.63	26.72	24.89	23.02
	FID ↓	5.21	7.84	11.26	15.03	18.92
	LPIPS ↓	0.092	0.118	0.151	0.186	0.221
W-mix	SSIM ↑	0.956	0.921	0.892	0.853	0.817
	PSNR ↑	32.18	29.27	27.41	25.43	23.78
	FID ↓	4.73	6.92	9.83	13.74	17.65
	LPIPS ↓	0.081	0.104	0.137	0.169	0.203
Both	SSIM ↑	**0.972**	**0.929**	**0.901**	**0.863**	**0.835**
	PSNR ↑	**34.81**	**30.71**	**28.43**	**26.10**	**24.36**
	FID ↓	**2.91**	**5.23**	**7.86**	**10.41**	**12.80**
	LPIPS ↓	**0.048**	**0.063**	**0.104**	**0.118**	**0.145**

2) Effect of Loss Function Design: Table VI evaluates different loss combinations across mask ratios. Using only L_1 and L_2 yields limited performance, especially for large masks. Adding LPIPS loss improves SSIM and PSNR while reducing perceptual errors, and including MSE loss further boosts FID under medium to high masks. The full combination achieves the best overall results, demonstrating the effectiveness of multi-loss joint optimization for both structural and perceptual restoration.

TABLE VI
ABLATION STUDY ON LOSS FUNCTION DESIGN

Loss	Metrics	10–20%	20–30%	30–40%	40–50%	50–60%
$L_1 + L_2$	SSIM ↑	0.912	0.885	0.852	0.808	0.762
	PSNR ↑	27.84	26.02	24.21	22.08	20.31
	FID ↓	9.86	13.24	18.01	23.92	29.75
	LPIPS ↓	0.158	0.182	0.214	0.253	0.298
$L_1 + L_2$+ LPIPS	SSIM ↑	0.941	0.907	0.876	0.835	0.792
	PSNR ↑	30.96	28.12	26.31	24.26	22.47
	FID ↓	6.12	8.94	12.86	17.41	21.93
	LPIPS ↓	0.108	0.132	0.165	0.203	0.241
$L_1 + L_2$+ MSE	SSIM ↑	0.949	0.915	0.885	0.845	0.806
	PSNR ↑	31.62	28.87	26.98	24.93	23.18
	FID ↓	5.64	8.21	11.73	15.86	20.12
	LPIPS ↓	0.096	0.121	0.154	0.191	0.227
LPIPS + MSE	SSIM ↑	0.965	0.924	0.896	0.858	0.823
	PSNR ↑	33.74	30.01	27.92	25.86	24.08
	FID ↓	3.84	6.37	9.42	13.06	16.91
	LPIPS ↓	0.062	0.085	0.118	0.151	0.183
Full	SSIM ↑	**0.972**	**0.929**	**0.901**	**0.863**	**0.835**
	PSNR ↑	**34.81**	**30.71**	**28.43**	**26.10**	**24.36**
	FID ↓	**2.91**	**5.23**	**7.86**	**10.41**	**12.80**
	LPIPS ↓	**0.048**	**0.063**	**0.104**	**0.118**	**0.145**

E. Other Studies

1) Object Removal Experiment: To further evaluate the effectiveness of the proposed method, object removal experiments are conducted on natural images using manually designed irregular masks. Compared with random masking, object removal introduces more complex missing regions with irregular boundaries and stronger semantic dependencies, posing greater challenges for structural modeling and detail reconstruction. As shown in Fig. 7, the proposed method effectively removes target objects while plausibly reconstructing background structures and textures, achieving consistent global layouts, smooth boundary transitions, and coherent local details. These results demonstrate the method's capability to restore missing regions in a semantically reasonable manner while reducing visual artifacts.

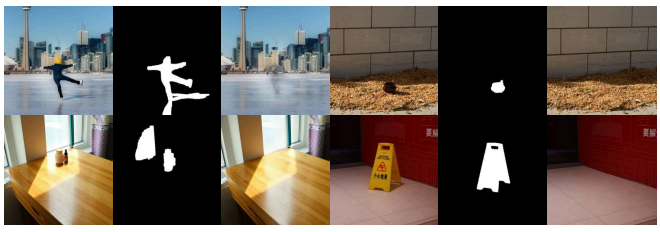

Fig. 7. Object removal result.

2) User Study: To assess perceptual quality, we conducted a user study on CelebA, Places2, and Paris Street View with varying mask ratios, presenting results from EC, HyperStyle [24], and PGI in random order. Participants scored images from 0 to 10 based on visual realism, structural plausibility, and texture naturalness, with averaged scores reported in Table VII. As expected, scores decreased with larger masks, reflecting the increased difficulty of inpainting. Nevertheless, our method consistently achieved higher scores than others, demonstrating its superior ability to preserve structure, generate plausible textures, and maintain overall perceptual quality.

TABLE VII
USER STUDY RESULTS

Dataset	Mask Ratio	EC	HyperStyle	Ours
CelebA	10–20%	5.4	7.9	**8.3**
	20–30%	5.1	7.4	**8.0**
	30–40%	4.9	6.9	**7.7**
	40–50%	4.6	6.5	**7.4**
	50–60%	4.3	6.2	**7.1**
Places2	10–20%	5.1	7.5	**8.1**
	20–30%	4.8	7.3	**7.8**
	30–40%	4.5	6.8	**7.4**
	40–50%	4.2	6.1	**7.1**
	50–60%	3.9	5.8	**6.8**
Paris Street View	10–20%	5.6	7.7	**8.4**
	20–30%	5.3	7.5	**8.1**
	30–40%	5.0	6.9	**7.8**
	40–50%	4.7	6.6	**7.5**
	50–60%	4.4	6.3	**7.2**

V. CONCLUSION

In this work, we present PGI, a StyleGAN inversion–based model for image inpainting. Experiments on the CelebA, Place2, and Paris Street View datasets demonstrate that PGI can generate structurally coherent and visually realistic results. These results indicate that the proposed method is effective in improving inpainting quality, particularly in preserving both structural integrity and texture details.

The proposed method still has certain limitations. Its performance depends on the representation capacity of the pre-trained generative prior, and extremely complex or ambiguous missing regions may lead to suboptimal inversion results. In addition, irregular mask patterns and out-of-distribution image content may pose challenges to the latent mixing strategy.

A detailed analysis of computational complexity will be explored in future work. We will also focus on improving the robustness of the proposed framework and investigating its extension to other related visual tasks, such as image editing.

REFERENCES

[1] J. C. Santos, T. P. A. H. Tomás, M. S. Seoane Santos, *et al.*, "The role of deep learning in medical image inpainting: A systematic review," *ACM Transactions on Computing for Healthcare*, vol. 6, no. 3, pp. 1–24, 2025.

[2] Q. Wang, S. He, M. Su, *et al.*, "Image inpainting methods: A review of deep learning approaches," *Symmetry*, vol. 18, no. 1, p. 94, 2026.

[3] K. Karwowska, D. Wierzbicki, and M. Kedzierski, "Image inpainting and digital camouflage: Methods, applications, and perspectives for remote sensing," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, 2025.

[4] X. Li, Y. Ren, X. Jin, *et al.*, "Diffusion models for image restoration and enhancement: A comprehensive survey," *Int. J. Comput. Vis.*, vol. 133, no. 11, pp. 8078–8108, 2025.

[5] X. Cao, "A review of image inpainting methods based on deep learning," in *Proc. 6th Int. Conf. Comput. Vis., Image Deep Learn. (CVIDL)*, 2025, pp. 356–360.

[6] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 2149–2159.

[7] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, "Large-scale image completion via co-modulated generative adversarial networks," arXiv:2103.10428, 2021.

[8] H. Zheng, Z. Lin, J. Lu, S. Cohen, E. Shechtman, C. Barnes, J. Zhang, N. Xu, S. Amirghodsi, and J. Luo, "Image inpainting with cascaded modulation GAN and object-aware training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Springer, 2022, pp. 277–296.

[9] L. Zhang, Y. Yu, J. Yao, and H. Fan, "High-fidelity image inpainting with multimodal guided GAN inversion," *Int. J. Comput. Vis.*, 2025.

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

[11] 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, 2017.

[12] 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, 2015.

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

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

[15] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, D. Ashukha, S. Silvestrov, and E. Burnaev, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2022, pp. 2149–2159.

[16] A. Lugmayr, M. Danelljan, A. Romero, and R. Timofte, "RePaint: Inpainting using denoising diffusion probabilistic models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11461–11471.

[17] Y. Poirier-Ginter and J.-F. Lalonde, "Robust unsupervised StyleGAN image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 22292–22301, 2023.

[18] W. Huang, Y. Deng, S. Hui, Z. Yu, and X. Xie, "Sparse self-attention transformer for image inpainting," *Pattern Recognit.*, vol. 145, p. 109897, 2024.

[19] S. Chen, A. Atapour-Abarghouei, and H. P. H. Shum, "HINT: High-quality inpainting transformer with mask-aware encoding and enhanced attention," *IEEE Trans. Multimedia*, vol. 26, pp. 7649–7660, 2024.

[20] 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.

[21] A. C. Bovik, *Handbook of Image and Video Processing*, 2nd ed. New York, NY, USA: Academic Press, 2005.

[22] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 6626–6637.

[23] 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 Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 586–595.

[24] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. H. Bermano, "HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, arXiv:2111.15666 [cs.CV], https://arxiv.org/abs/2111.15666.