Evaluating AI-based Image Inpainting Techniques for Facial Components Restoration Using Semantic Masks

Hussein Sharadga[a,b], Abdullah Hayajneh[c], Erchin Serpedin[c]

[a]*Texas A&M International University, 5201 University Blvd, Laredo, 78041, TX, USA*
[b]*The University of Texas at Austin, 2515 Speedway, Austin, 78712, TX, USA*
[c]*Texas A&M University, 400 Bizzell St, College Station, 77840, TX, USA*

Abstract

This paper presents a comparative analysis of advanced AI-based techniques for human face inpainting using semantic masks. The primary objective of this research is to assess the effectiveness of image inpainting methods in semantically restoring different facial elements. Our study demonstrates that image inpainting models experience significant challenges in reconstructing complete facial components. Unlike random masks, which often reveal parts of the main facial components (e.g., sections of the nose or mouth), semantic masks fully obscure them, potentially posing a greater challenge for inpainting methods. We evaluate the performance of various methods, including generative adversarial networks (GANs), transformers, and diffusion models, based on their ability to restore entire facial components.

To address these challenges and to enhance inpainting performance, we conduct three retraining processes using semantic masks, random masks, and a combination of both. This combined approach leverages the strengths of both mask types, enhancing the model's context-awareness and resulting in more realistic and accurate facial reconstructions.

Keywords: Image Inpainting, Semantic Masks, Face Restoration

1. Introduction

1.1. Motivation

Image inpainting is a sophisticated process for reconstructing missing or damaged areas of an image (1). It also involves seamlessly removing unwanted elements (such as snow (2) or shadows (3; 4)). It is essential in fields such as photography, film production, and digital art, where preserving the integrity of visual content is critical. By enabling the restoration of images to their original state or enhancing them for aesthetic appeal (5; 6), image inpainting plays a vital role in applications such as object removal and image restoration.

Recent advances in AI have significantly improved inpainting techniques, surpassing traditional methods that rely on local pixel information (7). The AI models offer greater accuracy and realism, particularly in reconstructing human faces, which require high precision and structural integrity. This paper focuses on the challenging task of reconstructing human faces using semantic masks that fully obscure key facial features, such as eyes, mouth, and hair. Unlike random masks, which may leave key facial components partially visible, semantic masks may demand more complex inpainting solutions.

1.2. Literature Review

Traditional inpainting methods rely on information from surrounding pixels but often fall short in capturing complex features (7). Deep learning significantly enhances inpainting algorithms by enabling more accurate, detailed, and realistic reconstruction of missing regions. Models integrating convolutional operations (8) and attention mechanisms (9) effectively capture both fine-grained textures and high-level semantic features, thereby improving the accuracy of predictions in damaged regions. The encoder-decoder Convolutional Neural Network (CNN) model (10) was trained to fill in the missing parts of an image with structures and textures that fit the existing image content. Generative Adversarial Networks (GANs) have further advanced inpainting by introducing adversarial training, which promotes

Email address: hssharadga@utexas.edu (Hussein Sharadga) more realistic and contextually accurate inpainted results (11).

Enhanced GAN architectures have been developed to improve inpainting quality and flexibility in (12; 13; 14). LaMa (12) is a new network architecture that uses Fourier convolutions with an image-wide receptive field and thus supports a large training mask. In (13), a GAN with gated convolution was trained to perform inpainting using free-form masks, which can be either freely drawn by a user or generated automatically. This flexibility facilitates more natural and context-aware image editing. In (14), an enhanced GAN-based model is proposed for high-resolution image inpainting, based on aggregated contextual transformations.

Transformers have become a powerful choice for image inpainting (15; 16; 17). The T-former model (15) introduces a transformer-based approach with an efficient attention mechanism that reduces computational complexity, effectively addressing CNN limitations like local priors and fixed spatial parameters through resolution-dependent attention. Another transformer-based model (16) handles large image holes by integrating both transformer and convolutional features effectively. The Continuous-Mask-Aware Transformer (CMT) (17) introduces a continuous mask to capture token errors, improving masked self-attention with overlapping tokens and refining inpainting results iteratively.

More recently, diffusion models—an alternative class of generative models—have been employed in image inpainting tasks (18; 19; 20; 21; 22). These models can be categorized as either preconditioned, which offer fast inference but are expensive to train, or postconditioned, which require no additional training but are computationally slower. LatentPaint (21) bridges these two paradigms by employing forward-backward fusion in a latent space, enhanced with a novel propagation module. Similarly, Latent Diffusion Models (LDMs) proposed in (19) exploit the latent space of powerful pretrained autoencoders, enabling high-resolution synthesis while reducing computational overhead.

1.3. Applications of Semantic Masks

Semantic masks, which completely obscure specific facial components, may present a more challenging task for image inpainting compared to random masks that leave parts of the main components of the face visible. Semantic masks require more structured and context-aware restoration. In facial recognition, semantic masks are used to protect privacy by masking certain facial features while preserving key identity markers (23; 24). Additionally, in medical imaging, they are employed to reconstruct facial features in cases of trauma or surgery, facilitating surgical planning and recovery visualization (25; 26). In creative industries, semantic masks enable accurate restoration or modification of facial features for digital art and content creation (27). These diverse applications highlight the broad utility of semantic masks in improving image restoration techniques across multiple domains.

1.4. Summary of Contributions

- Performance Analysis of Pre-trained Models on Semantic Masks:

  This paper evaluates the state-of-the-art image inpainting methods for reconstructing human faces using semantic masks. We assess these methods' capabilities to restore the main components of the human face, rather than using random masks—a focus that has not been studied before. Unlike random masking, which may leave parts of the main components visible, semantic masking completely obscures them, potentially posing a greater challenge for inpainting methods.

  Restoring human faces with semantic masks involves addressing the distinct challenges posed by each facial component. For example, hair requires modeling texture, flow, and color continuity; eyes demand symmetry, sharpness, and precise reconstruction of the iris and eyelashes; and the mouth involves the complex dynamics of teeth, lips, and proper alignment with the jawline.

- Retraining Models to Improve Performance:

  Additionally, we investigate the impact of different masking strategies on inpainting performance. To improve inpainting accuracy, we conduct three retraining processes using semantic masks, random masks, and a combination of both. The combined approach is proposed to improve the model's contextual awareness and inpainting performance.

1.5. Paper Layout

Section 2 describes the benchmark setup, including model selection, dataset, computing machine, and evaluation metrics. In Section 3, the selected models are compared at different resolutions (Section 3.1), and their limitations are highlighted (Section 3.2). Section 4 focuses on retraining the best-performing model using various masking strategies: random, semantic, and mixed masks. Future research directions are discussed in Section 5, followed by conclusion remarks in Section 6.

2. Benchmark Setup

2.1. Model Selection from Candidate Pool

The methods considered for comparison are listed in Table 1. According to MAT (16), four methods—MAT, LaMa, Co-Mod-GAN, and MADF—were identified as the most effective for face inpainting among nine evaluated techniques (AOT-GAN, MAT, LaMa, Co-Mod-GAN, MADF, ICT, HiFill, DeepFill v2, EdgeConnect) (14; 16; 12; 28; 29; 30; 31; 32; 33). Among these, MAT was found to be the most effective for inpainting faces with both small and large masks, as demonstrated in (16). In (16), these methods were tested using free-form masks generated by sampling rectangles and brush strokes of random shapes, sizes and locations. In contrast, this study evaluates these methods using semantic masks to specifically assess their ability to restore key facial components. Therefore, among these nine methods, MAT, LaMa, Co-Mod-GAN, and MADF are adapted in the current work.

However, a newer model, CMT (17), was found to outperform MAT in facial inpainting, followed by MAT, in a comparison of five methods (PIC (39; 38), ICT, BAT (37), MAT, and CMT) in (17). Thus, CMT is also adopted in the current work. Two other newer models with great potential are considered in this work. The first approach, MI-GAN (35), is a lightweight model primarily designed to run on mobile devices, yet it achieves performance comparable to state-of-the-art inpainting methods (MAT, LaMa, HiFill, Co-Mod-GAN, SH-GAN (36), ZITS (40), LDM (19)). A pluralistic image inpainting model with large masks represents another new model (34). This model is based on discrete latent codes and is referred to as “Latent or Latent-based” in this paper. Latent-based model (34) outperforms MAT, LaMa, MaskGIT (41), and PIC (39), as demonstrated in (34).

MAT (16) includes two models CelebA and FFHQ-512, both of which are evaluated in this study. The MAT CelebA model comes in two resolutions: 512x512 and 256x256, and both were tested. In this study, we used the base model of LaMa, as the authors (12) found that it outperformed other LaMa models on wide masks and did not significantly impact performance on narrow masks. LaMa was adapted to accept images and masks of any resolution (12). Since the CelebA dataset has a resolution of 1024x1024, we initially used this higher resolution but found LaMa to be ineffective at such high resolutions. Consequently, we evaluated LaMa at resolutions of 512x512 and 256x256. Co-Mod-GAN presents two versions: one implemented in an older version of TensorFlow and one in PyTorch. We used the PyTorch version for compatibility.Co-Mod-GAN includes two models with resolutions of 512x512 and 1024x1024, both of which were evaluated in this study.

RePaint (18), the diffusion-based model, was initially considered but ultimately evaluated on only 10 images due to its computational inefficiency: it requires about 10 hours to inpaint 10 images with 12 semantic mask classes (see mask classes in Table 3), with each class representing a different part of the face. RePaint takes about five minutes to inpaint a single image with one semantic mask class, making it impractical for large-scale testing.

In contrast, the other methods evaluated in this study, including MI-GAN, CMT, MAT, LaMa, Co-Mod-GAN, and MADF, can inpaint approximately 2,000 images within five minutes for a single semantic mask class, making them much more suitable for the scope and scale of this comparison. Latent-based model takes 10 minutes to inpaint about 2,000 images for a single semantic mask class, which is still acceptable.

2.2. Dataset

This study utilizes the CelebA dataset (42), which comprises about 30,000 high-quality images at a resolution of 1024×1024. The dataset provides a broad range of diversity in age, ethnicity, and background, as well as a variety of accessories like eyeglasses, sunglasses, and hats. It also includes detailed annotations for facial components, making it well-suited for training and evaluating face inpainting models with semantic masks. Some models require specific resolutions, so the images are rescaled accordingly to ensure compatibility.

2.3. Computing Resources

The following table outlines the specifications of the computing machine used for the experiments. It includes details on the GPU model, memory used, and other relevant components that contributed to the performance of the system during the testing phase. Efficient hardware resources played a crucial role in accelerating the training and inference processes, especially when working with high-resolution images and complex models. Additionally, the computational capacity allowed for multiple retraining scenarios and extensive evaluation of different inpainting methods.

2.4. Masks Generation

Various mask generation policies have been proposed including narrow masks, wide masks, box masks, box-and-narrow combination masks, and free-form masks

Method	Year	Resolution(s)	Model(s)
Latent (34)	2024	256	CelebA
MI-GAN (35)	2023	256	FFHQ
CMT (17)	2023	256	CelebA
SH-GAN (36)	2023	512	FFHQ
MAT (16)	2022	512, 256	CelebA-256/512, FFHQ-512
RePaint (18)	2022	256	CelebA
ICT (30)	2022	256	FFHQ
AOT-GAN (14)	2022	512	CelebA
LaMa (12)	2021	1024, 512, 256	CelebA-base
Co-Mod-GAN (28)	2021	1024, 512	FFHQ
MADF (29)	2021	512	CelebA
BAT (37)	2021	256	CelebA
PIC (38)	2021	256	CelebA
HiFill (31)	2020	512	CelebA
DeepFill (32)	2019	256	CelebA
EdgeConnect (33)	2019	256	CelebA

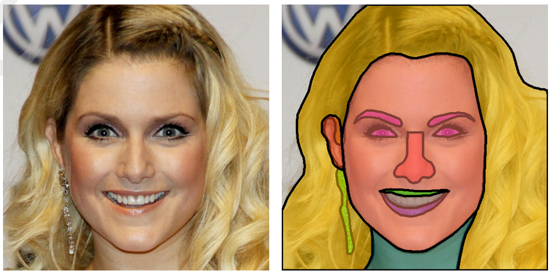
Table 1: Overview of different image inpainting methods, including publication year, resolutions, and corresponding model information. The methods selected for comparison are highlighted in bold: Latent-based, MI-GAN, CMT, MAT, RePaint, LaMa, Co-Mod-GAN, and MADF. Note: A total of 13 result extractions were performed, as some methods were evaluated across multiple resolutions and models.

Component	Specification
GPU Model	NVIDIA A800 Active
GPU Memory	40 GB per GPU
Number of GPUs	2
Driver Version	525.116.04
CUDA Version	12.0
CPU	Intel Xeon w9-3495X
CPU Cores	56 Cores, 1.9-4.8 GHz
RAM	256 GB (8 × 32 GB) 4800 MT/s
Total Storage Usage	286 GB

Table 2: System specifications of the computing machine used in the experiments.

(12; 13). The method of mask generation significantly influences overall model performance (12).

In this study, we evaluate the performance of different inpainting models using semantic masks. The dataset used in (43) includes images with corresponding semantic segmentations based on a 19-class model, where each pixel in the image is represented by an integer value from 0 to 18, corresponding to specific classes. Additionally, a more detailed segmentation is provided by a 34-class model from (43), which captures additional regions such as the cheeks, eyelids, forehead, jaw, and chin. However, the availability of labeled data for the 34-class model is limited (43; 44).



(a) Original image (b) Overlaid segmentation

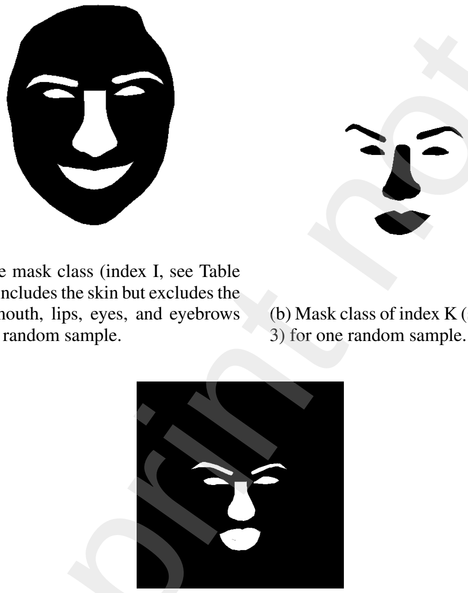
Figure 1: Original image with overlaid segmentation from CelebA dataset (42). The mouth region is divided into three segments: upper lip, mouth, and lower lip. The eyes, eyebrows, ears, hair, neck, skin, nose, and earrings are also segmented. An example of the CelebA 19-class model for face semantic segmentation used in the current study.

In this study, we used the 19-class model for face semantic segmentation with the CelebA dataset from (42) (see an example in Figure 1). Each image in (42) is accompanied by 19 corresponding mask images in black and white. The CelebA dataset provides segmentations for various facial regions, including the eyes, eyebrows, nose, upper lip, mouth, lower lip, hair, neck, ears, eyeglasses, skin, clothing, necklace, earrings, and hat.

We generate 12 mask classes, as shown in Table 3. The mask either obscures one component or a combination of components. Figure 2 shows the masks for classes (I, K, and L).

Mask Index	Face	Eyes	Nose	Lip (L)	Lip (U)	Mouth	Hair	Ears
A	×	✓	×	×	×	×	×	×
B	×	×	✓	×	×	×	×	×
C	×	×	×	✓	✓	✓	×	×
D	×	×	×	✓	×	×	×	×
E	×	×	×	×	✓	×	×	×
F	×	×	×	×	×	✓	×	×
G	×	×	×	×	×	×	✓	×
H	×	×	×	×	×	×	×	✓
I	✓	×	×	×	×	×	×	×
J	×	×	✓	✓	✓	✓	×	×
K	×	✓	✓	✓	✓	✓	×	×
L	✓	×	×	×	×	×	✓	✓

Table 3: Mask class index and corresponding facial parts combinations. Note: "Ears" includes the right ear, left ear, and earring and "Eyes" includes the eyebrows. "Lip (L)" refers to the lower lip and "Lip (U)" refers to the upper lip. The face refers to the skin excluding the neck, eyes, nose, mouth, and ears. A "✓" indicates the masked parts.



(a) Face mask class (index I, see Table 3) that includes the skin but excludes the nose, mouth, lips, eyes, and eyebrows for one random sample.

(b) Mask class of index K (see Table 3) for one random sample.

(c) Mask class of index L for one random sample.

Figure 2: Combined mask classes showing (a) Face mask class (index I), (b) Mask class of index K, and (c) Mask class of index L for random samples. Note: the masked areas are in black.

2.5. Evaluation Metrics

To evaluate different inpainting methods, we use perceptual metrics (FID, P-IDS, and U-IDS). Metrics like PSNR and SSIM do not correlate well with human perception of image quality (16). Among the perceptual metrics, FID is the most commonly used, as demonstrated in (34; 16; 17; 12; 35; 28; 29).

FID (Fréchet Inception Distance): Measures the distance between real and generated image distributions in feature space. It is computed as follows:

$$\mathrm{FID}(p_r, p_g) = \left\|\mu_r - \mu_g\right\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right) \quad (1)$$

where p_r and p_g represent the real and generated image distributions, respectively. Additionally, μ_r and μ_g are the mean feature vectors for the real and generated images, while Σ_r and Σ_g denote the covariance matrices for these distributions.

P-IDS (Perceptual Inception Distance Score): A variant of FID used to assess perceptual similarity, designed to align more closely with human visual perception, is defined as follows:

$$\text{P-IDS}(p_r, p_g) = \frac{1}{N}\sum_{i=1}^{N}\left(\|f(x_i) - f(x_i')\|_2^2\right) \quad (2)$$

where $f(x)$ is the feature vector extracted from the Inception network for image x, and x_i and x_i' are the real and generated images, respectively. The metric averages the perceptual differences over all samples.

U-IDS (Unsupervised Inception Distance Score): Assesses the quality of generated images using an unsupervised approach with the Inception model, without the need for labeled data, and it is calculated as follows:

$$\text{U-IDS}(p_r, p_g) = \frac{1}{N} \sum_{i=1}^{N} \left(\|f_u(x_i) - f_u(x_i')\|_2^2 \right) \tag{3}$$

where $f_u(x)$ is the unsupervised feature vector extracted from the Inception network for image x, and x_i and x_i' represent the real and generated images, respectively. This metric evaluates the similarity between distributions of real and generated images in the feature space.

3. Comparative Study

In this section, we compare the performance of various image inpainting models at different resolutions across multiple mask classes. Our analysis reveals that models perform differently depending on both the type of mask used and the resolution of the input images. The results highlight the strengths and weaknesses of each model in handling specific facial features, such as eyes, mouth, and hair. The following subsections provide a detailed comparison based on three key evaluation metrics: FID, P-IDS, and U-IDS, followed by a discussion of the strengths and weaknesses of each model as observed across various test conditions.

3.1. Scores Comparison

Comparing the results of this study with those in (16), we observe that semantic masks are more challenging to inpaint than random masks, as reflected in the evaluation metrics (higher FID values for semantic masks). Semantic masks fully obscure key facial features, whereas random masks may leave parts of these features visible, facilitating the inpainting process.

The models selected for comparison operate at different resolutions, with some capable of handling multiple resolutions. They are categorized into three groups based on their resolution:

3.1.1. Resolution 256

The FID values for different methods at a low resolution of 256 across mask classes (A-L) are shown in Table 4. Based on these FID values, MI-GAN ranks among the top three performing methods in 10 mask classes, Latent-based in 9 mask classes, and MAT in 8 mask classes. MI-GAN is the top-performing method in 4 mask classes, while Latent-based leads in 6 mask classes. As a relatively small model primarily designed for mobile devices, MI-GAN performs well. The changes in FID values across different mask classes and methods are consistent.

MI-GAN achieves the highest P-IDS value across all mask classes (see Table 4). MI-GAN, LaMa, and MAT generally emerge as the top three performing methods across different mask classes based on P-IDS values. For U-IDS values (see Table 4), MI-GAN achieves the highest U-IDS in 10 mask classes. MI-GAN, Latent-based, and MAT are generally the top three performing methods across different mask classes based on U-IDS values.

3.1.2. Resolution 512

The FID, P-IDS, and U-IDS values for different methods at a resolution of 512 across various mask classes are shown in Table 4. Based on FID, MAT with the FFHQ model ranks among the top three performing methods across 12 mask classes, being the best-performing method in 7 of them. MAT with the CelebA model ranks among the top three performing methods across 11 mask classes, while the MADF model ranks among the top three performing methods across 9 mask classes. However, Co-Mod-GAN generally outperforms MADF in both P-IDS and U-IDS.

Figure 3 illustrates the performance of different models in restoring facial key components at a resolution of 512. The figure highlights some limitations of these models. For example, in eye inpainting, LaMa and MADF generate blurry and mixed eyes. For nose masks, LaMa generates a small nose and a dual mouth. With face masks, MAT increases the thickness of the eyelashes; LaMa generates shorter face and face reflections; Co-Mod-GAN produces visible borders around the nose; and MADF generates an additional eye, overriding the existing one. For hair masks, MADF produces hair that lacks a realistic pattern. For mouth and nose masks, MADF generates teeth that override the lips.

3.1.3. Resolution 1024

Only two models, Co-Mod-GAN and LaMa, were trained to handle high-resolution images. The values for FID, P-IDS, and U-IDS are presented in Table 5. Co-Mod-GAN outperforms the LaMa model in 11 out of the 12 mask classes. The performance of Co-Mod-GAN and LaMa with high-resolution inpainting is illustrated in Figures 4 and 5. This demonstrates the capability of Co-Mod-GAN to better adapt to complex inpainting tasks at higher resolutions.

	256 Resolution					512 Resolution				
					FID					
Idx	**MI-GAN**	**Latent**	**MAT**	**CMT**	**LaMa**	**MAT**	**MAT$_1$**	**LaMa**	**Co-Mod-GAN**	**MADF**
A	1.175	1.685	1.800	2.098	1.782	1.141	1.206	5.965	1.513	2.204
B	1.086	1.407	1.386	1.376	0.951	0.734	0.661	2.147	1.103	0.734
C	2.659	1.744	2.610	2.677	3.729	2.030	1.370	3.682	2.261	1.425
D	0.917	1.327	1.414	1.230	1.202	0.661	0.601	1.224	0.914	0.589
E	1.276	1.154	1.256	1.239	1.396	0.770	0.648	1.313	0.807	0.625
F	1.734	2.146	2.793	2.877	2.053	2.054	1.375	3.503	1.629	1.166
G	10.536	8.541	9.757	14.109	17.384	8.928	7.406	21.869	8.968	16.546
H	3.260	3.784	3.870	4.011	3.909	3.444	2.956	3.963	3.994	3.272
I	6.333	7.603	6.130	5.709	9.252	5.418	5.056	11.176	10.136	6.483
J	3.071	2.155	2.833	2.887	3.995	2.312	1.619	5.811	3.038	1.711
K	3.558	2.799	3.472	3.739	5.002	2.927	2.364	11.503	3.962	3.172
L	66.48	46.35	91.71	288.99	114.35	56.72	85.56	176.68	97.91	114.76
					P-IDS					
Idx	**MI-GAN**	**Latent**	**MAT**	**CMT**	**LaMa**	**MAT**	**MAT$_1$**	**LaMa**	**Co-Mod-GAN**	**MADF**
A	2.46	0.75	0.60	0.25	0.40	5.47	3.71	0.10	4.87	0.50
B	3.25	0.80	1.25	0.75	1.85	9.65	11.55	0.65	7.50	6.10
C	2.51	1.30	0.85	1.10	0.65	9.92	9.62	0.90	9.17	6.27
D	2.01	1.20	1.25	1.76	1.51	9.28	10.04	4.62	8.08	6.57
E	1.60	1.00	1.05	1.00	1.40	7.12	8.68	4.86	6.62	4.81
F	0.25	0.00	0.00	0.00	0.08	1.39	2.13	0.25	2.29	1.06
G	0.45	0.00	0.00	0.00	0.00	1.06	3.49	0.00	0.66	0.00
H	1.38	0.07	0.21	0.14	0.21	3.53	5.33	1.66	1.38	2.70
I	0.15	0.00	0.00	0.00	0.00	0.65	2.55	0.00	0.00	0.00
J	2.51	1.05	1.05	0.70	0.35	9.07	10.88	0.05	6.52	5.56
K	1.85	0.35	0.25	0.10	0.05	4.00	3.55	0.00	2.15	0.50
L	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
					U-IDS					
Idx	**MI-GAN**	**Latent**	**MAT**	**CMT**	**LaMa**	**MAT**	**MAT$_1$**	**LaMa**	**Co-Mod-GAN**	**MADF**
A	9.66	4.94	4.82	2.11	3.01	12.73	10.82	0.13	11.02	1.98
B	14.10	7.87	9.35	8.70	11.72	22.92	26.15	2.72	18.53	20.43
C	8.45	9.35	6.52	6.27	2.48	20.28	22.96	2.96	19.47	16.84
D	13.75	9.23	8.18	9.71	11.57	22.96	25.77	15.13	23.28	21.20
E	12.46	10.53	9.60	10.43	9.68	23.72	25.58	15.15	21.49	23.14
F	0.57	0.25	0.04	0.00	0.20	2.45	4.62	0.94	3.92	4.17
G	1.36	0.00	0.10	0.00	0.00	3.08	7.23	0.00	1.11	0.03
H	3.15	0.41	0.73	0.66	0.83	7.33	12.66	4.50	4.05	7.12
I	0.70	0.00	0.15	0.15	0.00	2.92	7.07	0.00	0.00	0.02
J	9.00	7.22	5.51	5.01	1.90	19.65	20.85	0.10	14.24	14.11
K	6.17	2.92	2.20	0.72	0.25	9.40	9.20	0.00	6.57	1.50
L	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 4: FID, P-IDS, and U-IDS values for various methods across mask classes (A-L) at resolutions of 256 and 512. The three top-performing methods (i.e., those with the lowest FID values, those with the highest P-IDS values, those with the highest U-IDS values) for each mask class at each resolution are highlighted in blue. The 512-resolution MAT includes two models: CelebA (denoted as MAT) and FFHQ (denoted as MAT$_1$). Based on the FID, P-IDS, and U-IDS values, MAT$_1$ consistently ranks among the top three models for different mask classes.

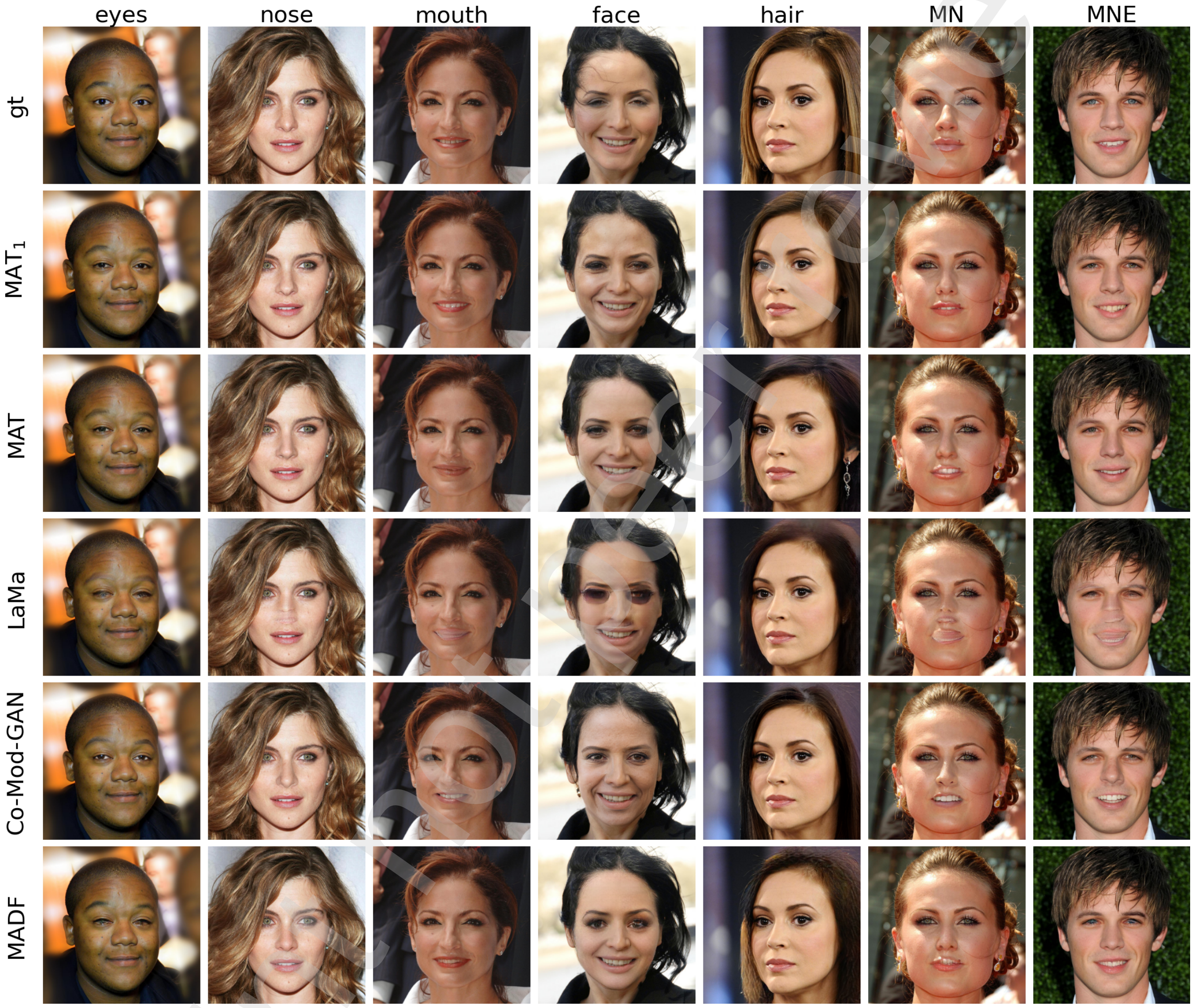


Figure 3: Restoration of key facial components at 512 × 512 resolution using various models. Some model limitations include: blurry or mixed eyes, a short nose, fused teeth, visible mask borders, blended lips, a shortened face, thick black eyelashes, unrealistic hair, and teeth overriding the lips. View better when zoomed in. The "gt" refers to the ground truth image, "MN" refers to the mouth and nose mask class (Class J in Table 3), and "MNE" refers to the mouth, nose, and eyes mask class (Class K in Table 3). The MAT method includes two models: CelebA (denoted as MAT) and FFHQ (denoted as MAT$_1$).

	FID		P-IDS		U-IDS	
Idx	**Co-Mod-GAN**	**LaMa**	**Co-Mod-GAN**	**LaMa**	**Co-Mod-GAN**	**LaMa**
A	2.223	5.904	2.41	0.15	4.87	0.20
B	0.947	2.299	8.35	0.35	18.45	1.15
C	2.818	5.075	9.57	0.00	16.42	0.03
D	0.764	1.031	11.94	5.67	24.28	14.43
E	0.722	0.369	13.24	18.91	26.91	31.02
F	1.837	5.363	3.84	0.74	6.21	1.19
G	11.597	41.848	0.40	0.00	1.21	0.00
H	3.838	3.864	3.11	2.21	6.50	5.22
I	11.235	33.611	0.00	0.00	0.00	0.00
J	3.934	9.202	3.21	0.00	7.99	0.00
K	6.23	18.583	0.00	0.00	0.38	0.00
L	65.73	206.07	0.00	0.00	0.00	0.00

Table 5: FID, P-IDS, and U-IDS values for (1) Co-Mod-GAN and (2) LaMa at a high resolution of 1024 are reported across mask classes (A-L). The top-performing method (i.e., the one with the lowest FID value, the highest P-IDS value, or the highest U-IDS value) for each mask class is highlighted in blue.

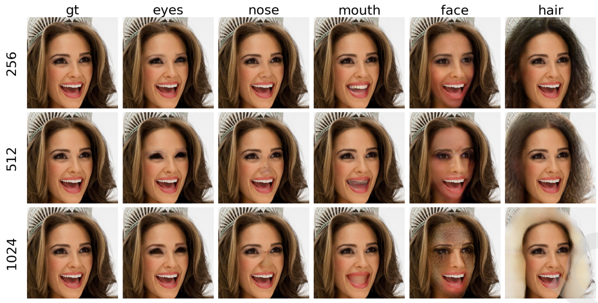


Figure 4: Restoring face key components at 256, 512, and 1024 resolutions using LaMa Model. The LaMa model faces challenges in restoring effectively at resolutions higher than 256. View better when zoomed in.

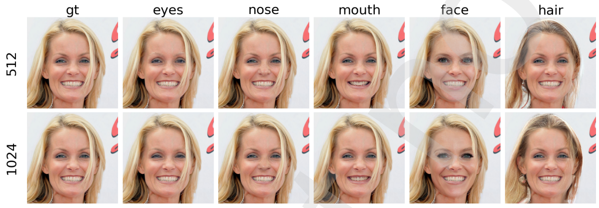


Figure 5: Restoring key facial components at 512 and 1024 resolutions using the Co-Mod-GAN model. The Co-Mod-GAN model performs effectively at high resolutions, such as 1024, and is comparable to its performance at 512. View better when zoomed in.

Co-Mod-GAN demonstrates reliable performance for high-quality image inpainting, while LaMa, though capable of handling various resolutions, does not scale well at higher resolutions. This results in Co-Mod-GAN being the more effective model for tasks requiring high-resolution image restoration.

3.2. *Findings and Limitations Across Models*

The following observations and limitations (summarized in Table 6) were identified for each method based on the analysis of all the generated images. Although some methods achieve relatively low FID values, they still have limitations and may not always perform optimally. This indicates that no single method can be considered superior in all scenarios, and there is still room for improvement. Overall, these observations suggest that ongoing research and refinement are needed to develop more versatile and reliable inpainting methods.

All inpainting models consume less than 0.3 seconds to inpaint one image, except for RePaint, which requires approximately 5 minutes. The computation time for each model is summarized in Table 7.

The limitations include various inconsistencies across facial components. For example, eyes may appear unrealistic with issues like thick eyelashes, pupil misalignment, or full black eyes in some models. The mouth may exhibit problems such as merged lips, unrealistic teeth, or misaligned lower lips. Skin inconsistencies include visible mask borders, skin tones that do not blend well, and sometimes unnatural reflections or dark skin patches. Hair restoration often suffers from unrealistic textures, color continuity problems, or incomplete coverage, especially at higher resolutions. Ears may be missing, incomplete, or replaced by hair in some cases. Additionally, noses, while generally well restored in most models, sometimes exhibit incomplete or unrealistic results. These imperfections can significantly impact the overall realism of the generated images. Overall, while these models generally perform well in specific tasks, such limitations indicate there is still room for improvement in generating more accurate and seamless facial inpainting.

Table 6: Common limitations and observations for different inpainting models.

Model	Details
1. MAT	**1. Eye:** MAT may increase the size of the lacrimal caruncle and the thickness of the eyelashes. **2. Mouth:** MAT produces better results for full-mouth restoration (a mask of index C) rather than for a partial restoration. MAT reduces gummy smiles, and frequently generates plump lips. **3. Face:** MAT lightens skin tone, thins faces, and sometimes enlarges lips when not masked. **4. Ears:** May generate incomplete ears or only one ear visible while the other is replaced by hair. **5. Hair:** Hair often looks realistic in terms of flow and color continuity. **6. Nose:** Nose blends seamlessly with the skin and looks realistic.
2. Co-Mod-GAN	**1. Eye:** Eyes look realistic, but eyebrows may not blend seamlessly with the surrounding skin. **2. Mouth:** Lower lip color mismatch, visible borders, and unrealistic teeth. **3. Face:** Inpainted skin borders sometimes don't blend well, dual eyes may appear. **4. Ears:** Generates incomplete ears. **5. Hair:** May use masked area to enlarge the face, hair continuity is not always maintained. **6. Nose:** Works well, but borders of the inpainted nose may not blend seamlessly.
3. MADF	**1. Eye :** Eyes are often unrealistic, missing pupils, blurry, or mixed. **2. Mouth :** Teeth often appear unrealistic, lips may blend with teeth or appear blurry. **3. Face:** Dual eyes/eyebrows, and inpainted skin borders may not blend well. **4. Ears:** May generate incomplete ears. **5. Hair:** Hair appears unrealistic with unnatural continuity. **6. Nose:** Nose blends seamlessly and looks realistic.
4. CMT	**1. Eye :** Eyes may have thick eyelashes, light or gray eyebrows. **2. Mouth :** Lips may fuse into a single piece without separation; teeth may be unrealistic. **3. Face:** Skin sometimes appears dark or blackish. **4. Ears:** One ear missing or replaced with hair. **5. Hair:** Hair appears unrealistic, often in chunks. **6. Nose:** Nose may be incomplete.
5. Latent-based	**1. Eye :** Eyes may have black eyelashes. **2. Mouth :** Teeth may be mixed, cracked, or uneven. **3. Face:** Mask borders visible, lashes enlarged. **4. Ears:** Masked ear may be inpainted with hair. **5. Hair:** Hair often looks realistic. **6. Nose:** Nose looks realistic.
6. MI-GAN	**1. Eye :** Eyes may have thick black eyelashes, pupil position issues, or different colors. **2. Mouth :** Teeth may be connected, cracked, or uneven; lips may be swollen. **3. Face:** Sometimes enlarges lower lip or thickens eyelashes; borders visible. **4. Ears:** Ears may be incomplete or missing. **5. Hair:** Part of the hair mask may be inpainted with ear or background elements. **6. Nose:** Nose looks realistic.
7. LaMa	**1. Eye :** Eyes may be fully black with no pupils. **2. Mouth :** Teeth may not be restored; lips may appear with missing or small teeth. **3. Face:** May generate face reflections if the face is masked. **4. Ears:** Does not restore ears, replacing them with hair. **5. Hair:** Hair restoration ineffective, especially at higher resolutions. **6. Nose:** Nose restoration fails above resolution 256.
8. RePaint	**1. Eye :** Pupils may have different colors; very black eyelashes. **2. Mouth :** Lower lip may be bigger, with missing lower jaw teeth. **3. Face:** Face looks realistic. **4. Ears:** Ears may be incomplete. **5. Hair:** Hair looks realistic, but RePaint ignores fine details. For example, if the mask doesn't fully cover the hair, RePaint doesn't utilize the residual hair. **6. Nose:** Nose looks realistic.

4. Evaluation of Models Retrained with Semantic, Random, and Mixed Masks

The resolution of the images used for this evaluation is 512. Since MAT has demonstrated the greatest potential for inpainting images at this resolution, as shown in Table 4, it was chosen for the experiments described in this section.

In this evaluation, MAT was retrained using semantic masks. To compare the performance of MAT trained with semantic masks against MAT trained with random masks, a version of MAT was also retrained with random masks under the same configuration to ensure fairness. It is worth noting that the retrained MAT achieved FID, P-IDS, and U-IDS values close to those of the pretrained MAT model provided by its authors. These models were then evaluated for their ability to inpaint faces using semantic masks.

MAT was retrained using 24,000 images and tested on a dataset of 6,000 images. Each training session took approximately 7 days and 13 hours. The FID values achieved on semantic masks by training MAT with semantic masks are compared to the FID values achieved on semantic masks by MAT trained with random masks. It is shown that MAT trained with random masks outperforms MAT trained with semantic masks for inpainting faces with semantic masks

MAT trained with random masks alone can outperform MAT trained with semantic masks alone for inpainting tasks because random masks promote better generalization and reduce overfitting. Training with random masks exposes the model to a diverse range of missing regions, encouraging it to learn broader contextual relationships across the entire image. This results in a more robust model capable of handling varied inpainting scenarios. In contrast, semantic masks provide predefined regions, which can lead to overfitting of specific features and reduce the model's ability to generalize. As a result, MAT trained with random masks alone tends to produce more realistic inpainted images, achieving lower FID scores compared to MAT trained with semantic masks alone.

To leverage the strengths of both random and semantic masks and improve inpainting performance, we retrain the MAT model using a combination of both mask types. This combined approach, referred to as the "mixed" masking strategy, exposes the model to a diverse set of regions for inpainting. By randomly selecting between random and semantic masks for each image, the model is forced to learn a broader range of contextual relationships. This allows it to adapt to varying types of missing information, which is particularly beneficial for inpainting complex features such as facial components.

The inclusion of semantic masks provides the model with more structured guidance for inpainting, while random masks encourage flexibility and better generalization by presenting the model with more unpredictable scenarios. This hybrid strategy enhances the model's context-aware capabilities, leading to more realistic and accurate reconstructions of facial features, especially in regions where fine details are crucial. The model trained with mixed masks achieves a lower FID value across almost all mask indices, as shown in Table 8. This suggests that combining the two masking strategies allows the model to outperform its counterparts trained with only random or only semantic masks, particularly for inpainting images with semantic masks. The mixed mask approach, therefore, strikes an optimal balance between flexibility and structure, enabling the model to handle a wider variety of inpainting tasks with greater fidelity and realism.

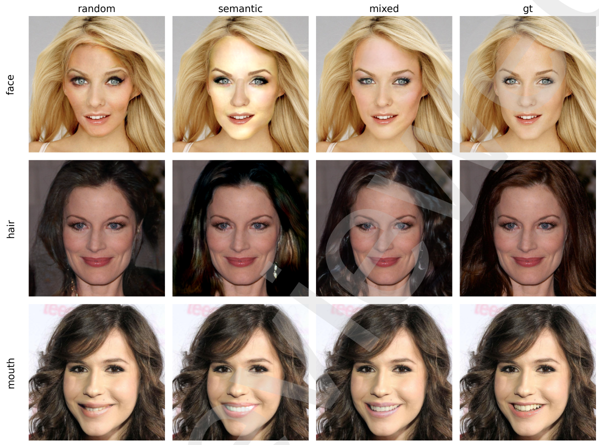


Figure 6: Restoring key facial components using 512-resolution MAT models trained with random, semantic, and mixed masks. Training with mixed masks enhances inpainting performance. View better when zoomed in.

Based on Table 8, there is a notable difference between the FID values for MAT trained with random, semantic, and mixed masks for the mask classes with indices (C, G, and I), corresponding to full mouth, hair, and face masks. To further illustrate these differences, we randomly select samples and visualize them in Figure 6.

For the face mask, as shown in Figure 6, MAT trained with random masks achieves a blended color, but the face appears short and inconsistent. MAT trained with semantic masks results in a face size and shape closer to the ground truth; however, it does not blend well with the surrounding color. The face exhibits a noticeably different color tone compared to the nose. MAT trained with mixed masks achieves better results than MAT trained with only random masks or only semantic masks.

For the hair mask, MAT trained with random masks may struggle with hair flow, while MAT trained with a semantic mask may face challenges with color continuity. Using MAT trained with mixed masks achieves a better balance between hair flow and color consistency, as shown in Figure 6.

For the mouth mask, MAT trained with mixed masks may outperform both MAT trained with a semantic mask and MAT trained with a random mask, as shown in Figure 6.

5. Future Work

A promising direction for future work is the development of specialized AI models that focus on individual facial compo-

Table 7: Performance summary for various models

Model	Performance
MAT	Inpainting 2,000 images takes approximately 5 minutes at a resolution of 512, and about 3 minutes to inpaint 2,000 images at a resolution of 256.
Co-Mod-GAN	Inpainting 2,000 images takes 1 minute at resolution of 512 and 2 minutes at resolution of 1024.
MADF	Inpainting 2,000 images takes approximately 6 minutes.
CMT	Inpainting 2,000 images takes about 1 minute.
Latent-based	Inpainting 2,000 images takes about 10 minutes.
MI-GAN	Inpainting 2,000 images takes about 2 minutes.
LaMa	Inpainting 2,000 images takes 30–60 seconds at 256×256, 90 seconds at 512×512, and 5 minutes at 1024×1024.
RePaint	Inpainting takes about 5 minutes per image.

	Mask Type		
Idx	Random	Semantic	Mixed
A	1.4597	1.7819	1.2535
B	0.9326	0.9510	0.9280
C	2.0123	3.7290	1.5901
D	0.7546	1.2019	0.7056
E	0.7410	1.3962	0.5638
F	2.3592	2.0534	1.2433
G	12.5149	17.3838	7.1022
H	3.9033	3.9087	3.3346
I	8.1616	9.2519	5.6928
J	2.4613	3.9950	2.2000
K	3.5571	5.0020	3.0273
L	115.513	114.353	130.824

Table 8: Comparison of FID values for MAT trained with random, semantic, and mixed masks. "Mixed" refers to training with both random and semantic masks, where each image is randomly masked with either a random mask or a semantic mask. The two top-performing methods (i.e., those with the lowest FID values) for each mask class are highlighted in blue. The results show that MAT trained with mixed masks outperforms MAT trained with either only semantic masks or only random masks in terms of FID, particularly for inpainting images with semantic masks.

nents. For example, an AI model could be trained specifically for inpainting hair, which could potentially improve performance by allowing the model to learn more specialized patterns and structures associated with that particular feature.

Additionally, a combined approach could be proposed, where the model is trained using both random masks and masks focusing specifically on hair. This approach may help the AI learn not only the detailed characteristics of individual facial parts but also the relationships between different face components, ultimately improving the model's ability to perform more accurate and context-aware inpainting tasks.

6. Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT to improve the readability of the paper. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

7. Conclusion

In conclusion, while AI-based inpainting methods demonstrate promising capabilities in restoring key facial components with realistic and contextually appropriate results, there are still areas that require improvement. Challenges persist in achieving seamless blending, handling complex facial structures, and preserving finer details, such as those in the eyes, teeth, hair, and eyelashes. Despite these challenges, ongoing advancements in AI algorithms and deep learning techniques offer significant potential for overcoming these limitations in the near future.

Furthermore, while semantic masks present a significant challenge for key facial component restoration, they can enhance the AI model's contextual awareness and inpaiting performance when combined with random masks. This hybrid approach not only improves the inpainting process but also enables more accurate and natural facial feature restoration in diverse settings.

References

[1] R. Ashwini, R. Vani, and S. Suvitha. Wholeness unveiled: Pioneering the fusion of denoising and inpainting for mastery in image restoration. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–6, 2024.

[2] Bodong Cheng, Juncheng Li, Ying Chen, and Tieyong Zeng. Snow mask guided adaptive residual network for image snow removal. *Computer Vision and Image Understanding*, 236:103819, 2023.

[3] Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. Shadowformer: Global context helps image shadow removal. *arXiv preprint arXiv:2302.01650*, 2023.

[4] Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. Shadowdiffusion: When degradation prior meets diffusion model for shadow removal. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14049–14058, 2023.

[5] Jiacheng Zhu, Junjie Wen, Duanqin Hong, Zhanpeng Lin, and Wenxing Hong. Uir-es: An unsupervised underwater image restoration framework with equivariance and stein unbiased risk estimator. *Image and Vision Computing*, 151:105285, 2024.

[6] Xin Pan, Hao Zhai, You Yang, Lianhua Chen, and Anyu Li. Improving multi-focus image fusion through noisy image and feature difference network. *Image and Vision Computing*, 142:104891, 2024.

[7] Zishan Xu, Xiaofeng Zhang, Wei Chen, Minda Yao, Jueting Liu, Tingting Xu, and Zehua Wang. A review of image inpainting methods based on deep learning. *Applied Sciences*, 13(20), 2023.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[10] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, 2020.

[11] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[12] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor S. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182, 2021.

[13] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas Huang. Free-form image inpainting with gated convolution. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479, 2019.

[14] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 29(7):3266–3280, 2023.

[15] Ye Deng, Siqi Hui, Sanping Zhou, Deyu Meng, and Jinjun Wang. T-former: An efficient transformer for image inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6559–6568, New York, NY, USA, 2022. Association for Computing Machinery.

[16] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[17] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13169–13178, 2023.

[18] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11451–11461, 2022.

[19] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[20] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), July 2023.

[21] Ciprian Corneanu, Raghudeep Gadde, and Aleix M Martinez. Latentpaint: Image inpainting in latent space with diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4334–4343, 2024.

[22] Seung-Lee Lee, Minjae Kang, and Jong-Uk Hou. Localization of diffusion model-based inpainting through the inter-intra similarity of frequency features. *Image and Vision Computing*, 148:105138, 2024.

[23] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14994–15003, 2022.

[24] Yaofang Zhang, Yuchun Fang, Yiting Cao, and Jiahua Wu. Rbgan: Realistic-generation and balanced-utility gan for face deidentification. *Image and Vision Computing*, 141:104868, 2024.

[25] Abdullah Hayajneh, Erchin Serpedin, and Mitchell Stotland. Automatic semantic in-painting image normalization for facial anomaly appraisal. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pages 1501–1505, 2024.

[26] Abdullah Hayajneh, Erchin Serpedin, Mohammad Shaqfeh, Graeme Glass, and Mitchell A. Stotland. Cleftgan: Adapting a style-based generative adversarial network to create images depicting cleft lip deformity. *arXiv preprint arXiv:2310.07969*, 2023. Available online.

[27] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5103–5112, 2020.

[28] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[29] Manyu Zhu, Dongliang He, Xin Li, Chao Li, Fu Li, Xiao Liu, and Errui Ding. Image inpainting by end-to-end cascaded refinement with mask awareness. *IEEE Transactions on Image Processing*, 30:4855–4866, 2021.

[30] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4672–4681, 2021.

[31] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7505–7514, 2020.

[32] Jiahui Yu, Zhe L. Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Free-form image inpainting with gated convolution. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4470–4479, 2018.

[33] K Nazeri. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[34] Haiwei Chen and Yajie Zhao. Don't look into the dark: Latent codes for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7591–7600, 2024.

[35] Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, and Humphrey Shi. Mi-gan: A simple baseline for image inpainting on mobile devices. In *2023 IEEE/CVF International Conference*

on Computer Vision (ICCV), pages 7301–7311, 2023.

[36] Xingqian Xu, Shant Navasardyan, Vahram Tadevosyan, Andranik Sargsyan, Yadong Mu, and Humphrey Shi. Image completion with heterogeneously filtered spectral hints. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4580–4590, 2023.

[37] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[38] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic free-from image completion. *International Journal of Computer Vision*, pages 1–20, 2021.

[39] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[40] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11348–11358, 2022.

[41] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. Maskgit: Masked generative image transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11305–11315, 2022.

[42] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[43] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models, 2021.

[44] Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, and Sanja Fidler. Dreamteacher: Pretraining image backbones with deep generative models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16652–16662, 2023.