CSF-GAN: Cross-modal Semantic Fusion-based Generative Adversarial Network for Text-guided Image Inpainting

Shilin Zhang[1] , **Suixue Wang**[2] , **Qingchen Zhang**[3,*] , **Liang Zhao**[4,*] , **Weiliang Huo**[2] , **Sijia Hou**[4] , **Chunjiang Fu**[4]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of Information and Communication Engineering, Hainan University, Hainan, China
[3]School of Computer Science and Technology, Hainan University, Hainan, China
[4]School of Software, Dalian University of Technology, Dalian, China
zhang_shilin_sd@163.com, {wangsuixue, zhangqingchen}@hainanu.edu.cn, liangzhao@dlut.edu.cn, wlhuo@hainanu.edu.cn, {housj, fuchunjiang}@mail.dlut.edu.cn

Abstract

Most visual-guided image inpainting methods based on generative adversarial networks (GANs) struggle when the missing region has weak correlations with the surrounding visual context. Recently, diffusion-based methods guided by textual context have been proposed to address this limitation by leveraging additional semantic information to restore corrupted objects. However, these models typically involve more parameters and exhibit slower generation speeds compared to GAN-based approaches. To address this problem, we propose a novel text-guided image inpainting model, the cross-modal semantic fusion generative adversarial network (CSF-GAN). CSF-GAN is designed as a one-stage GAN with the following key contributions. First, a novel semantic fusion module (SFM) is introduced to integrate sentence- and word-level textual context into the inpainting process, enabling more effective guidance from multi-granularity semantic information. Second, a newly designed word-level local discriminator provides detailed feedback to the generator, enhancing the accuracy of generated content in alignment with word-level semantics. Third, two loss functions, the inpainting loss and edge loss, are employed to enhance both structural coherence and textural realism in the generated results. Extensive experiments on two benchmark datasets demonstrate that CSF-GAN outperforms state-of-the-art methods.

1 Introduction

Image inpainting aims to generate missing regions and reconstruct visually plausible images, posing a significant challenge in computer vision. It is crucial for various applications, including restoring damaged artwork, editing digital images, and removing unnecessary objects.

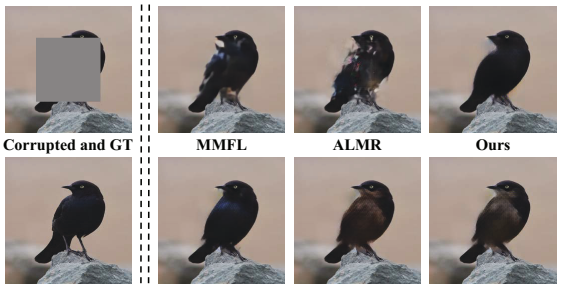

Figure 1: The first row presents the comparison of generated images between the proposed CSF-GAN and other advanced text-guided image inpainting methods, given the text **a**. The second row shows three examples of generated images based on CSF-GAN, with different texts as guidance (text **b**-**d**).

Generative adversarial network (GAN)-based methods are widely used for image inpainting [Iizuka *et al.*, 2017; Pathak *et al.*, 2016; Yin *et al.*, 2024; Zheng *et al.*, 2019], typically employing an encoder-decoder architecture for image synthesis and a discriminator to distinguish between generated and real images. However, their performance degrades when contextual information is weakly correlated with surrounding pixels. For example, in datasets like CUB-200-2011 and Oxford-102, where objects are centered, central masking makes inpainting more difficult than random masking.

Compared to internal guidance, providing inpainting models with external guidance offers a promising approach to controlling the output. In internal guidance, Feng *et al.* [Feng *et al.*, 2022]infer the content of corrupted regions using learned inter-image reasoning priors that capture semantic distribution patterns among similar images. However, text descriptions contain richer semantic information, making them more effective for guiding content generation in missing regions. As shown in Figure 1, using different text prompts results in distinct image content. Therefore, Lin et al. [Lin *et al.*, 2020] introduce a coarse-to-fine inpainting

*Corresponding authors: Qingchen Zhang and Liang Zhao.

model (MMFL, as shown in Figure 1) that captures essential information in the damaged areas. Furthermore, Wu et al. [Wu *et al.*, 2021] develop a module for reconstructing masks, ensuring that the restored object closely resembles the original object in the initial image (ALMR, as depicted in Figure 1). However, the two methods inject semantic information into the inpainting process by simply concentrating on hidden visual features and visual text features, leading to the insufficient utilization of semantic information. Besides, the primary loss in their loss functions is reconstruction loss, which can only guarantee accurate structures but fail to consider the high-resolution textures, leading to unsatisfying inpainting results. To tackle these problems, our method injects sentence-wise and word-wise semantic information into the inpainting process to modulate visual features, which can better employ text semantics to guide this inpainting task. Also, our method uses the inpainting loss and the edge loss to guarantee accurate structures and high-resolution textures in inpainting images. As shown in Figure 1, CSF-GAN generates realistic and natural outcomes in terms of structure and texture according to the text description, while prior methods [Lin *et al.*, 2020; Wu *et al.*, 2021] are unable to generate visually plausible results.

In recent years, diffusion-based models [Rombach *et al.*, 2022] have demonstrated notable performance in image inpainting. However, these models are not without limitations. Firstly, the number of parameters in diffusion-based models significantly exceeds that of GAN-based models, resulting in higher demands on the machine's GPU configuration. Secondly, diffusion-based models require numerous iterations to generate images, resulting in a generation speed that is considerably slower than that of GAN-based models. In practical applications, generation speed serves as a critical performance indicator. Consequently, we continue to utilize GAN as our foundational model.

This paper presents a cross-modal semantic fusion approach to guide the restoration of damaged image pixels. Prior research employs a two-stage model for image restoration, wherein a shallow generator first produces a coarse-grained image, which is subsequently used as input for a deeper generator to generate a fine-grained image. In contrast, we propose a one-stage restoration model with fewer parameters that directly generates fine-grained images. Furthermore, to effectively leverage semantics from textual descriptions, we introduce a semantic fusion module (SFM) to facilitate the fusion of coarse-to-fine features between textual and visual components. The primary contributions of this research are as follows:

- We first introduce a one-stage cross-modal semantic fusion generative adversarial network with fewer model parameters for image inpainting based on text guidance.
- To better leverage the semantics in text descriptions, we propose a semantic fusion module (SFM) to embed sentence-level context and word-level context into the restoration process.
- We develop a novel word-level local discriminator, specifically for discriminating the missing patches from the word-level semantics in an adversarial way.
- An inpainting loss and an edge loss are introduced to optimize the proposed network considering structural and textural information respectively.

2 Related Work

2.1 Image Inpainting

Image inpainting aims to reconstruct damaged regions and generate visually realistic images. In recent years, deep learning methods, specifically Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) [Du *et al.*, 2024; Chen *et al.*, 2022], have demonstrated significant efficacy in this domain. For instance, Pathak et al. [Pathak *et al.*, 2016] introduce an encoder-decoder framework that integrates adversarial loss to enhance content comprehension across the entire image. To establish semantic relevance between the missing and existing areas, Liu et al. [Liu *et al.*, 2019] propose a two-stage architecture that employs a novel coherent semantic attention mechanism.

2.2 Text-guided Image Generation and Manipulation

Text-to-image generation is to create images based on text descriptions. Extensive GAN-based approaches are explored about text-to-image generation and can obtain photo-realistic images that match text priors. Reed et al. [Reed *et al.*, 2016] are the first to demonstrate that GANs can generate images conditioned on human-written descriptions. Zhang et al. [Zhang *et al.*, 2017; Zhang *et al.*, 2019] decompose this complex problem into several sub-problems and stack multiple GANs to synthesize images of varying sizes gradually. Xu et al. [Xu *et al.*, 2018] introduce a word-level attention mechanism within GANs, enabling the generation of specific image sub-regions based on the most relevant words.

Different from text-to-image generation, text-guided image manipulation focuses on modifying specific visual features of an image based on textual input while preserving irrelevant parts of the image. In [Nam *et al.*, 2018], a text-adaptive discriminator is introduced to provide fine-grained feedback to the generator for the generation of specific visual content. Li et al. [Li *et al.*, 2020a] propose a text-image affine combination module that selects image regions corresponding to the input text and a detail restoration module that fulfills any missing content.

Compared with the text-to-image generation task, it is more strict for the requirement of the text-guided image inpainting task. This is because the inpainting result must align with both the text and the existing content. Unlike text-guided image manipulation, which utilizes the entire image as a reference, the inpainting task only employs the semantics of the text and the existing image regions to restore damaged areas, making it more challenging.

2.3 Text-guided Image Inpainting

The semantics in text play a crucial role in image inpainting, leading to increased interest in the text-guided image inpainting task. Wu et al. [Wu *et al.*, 2021] propose a mask recon-

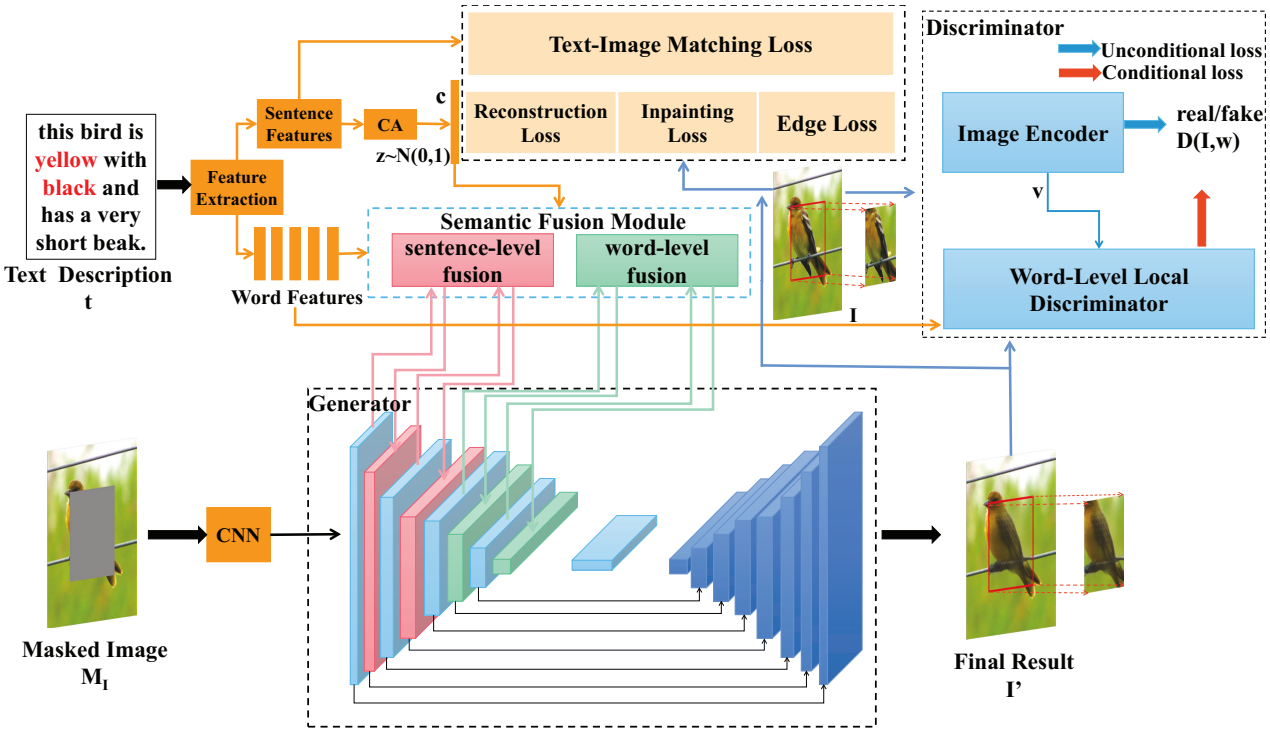

Figure 2: The structure of our generative adversarial network that utilizes cross-modal semantic fusion.

struction module to ensure that the primary object in the generated image closely resembles that in the source image. Li et al. [Li *et al.*, 2023] provide a visual-textual modalities fusion module to extract more valuable and informative textual features for image inpainting.

However, these methods treat text-guided image inpainting as a multi-stage or two-stage task, which is more complex than our one-stage model. Additionally, they primarily rely on reconstruction loss, which is insufficient for generating high-resolution textures in completed images.

3 Cross-Modal Semantic Fusion-based Generative Adversarial Network

Given a text description t and a masked image M_I, our method aims to produce an inpainting image I' which aligns with t and the existing image content. The structure of CSF-GAN is presented in Figure 2. It contains three components: the Generator for Image Inpainting, the Semantic Fusion Module, and the Word-Level Local Discriminator.

3.1 Generator for Image Inpainting

The input of the generator M_I is a $3 \times 256 \times 256$ image with a central hole, and the output I' is the painted image, matching the size of M_I. The generator is based on the UNet architecture [Ronneberger *et al.*, 2015], with both the encoder and the decoder consisting of 8 layers of neural networks. Before entering the encoder, the input image is first processed by a convolutional operation to increase the number of channels in the feature map to 64. In the encoder, each layer consists of two convolution blocks. To prevent convolution operation

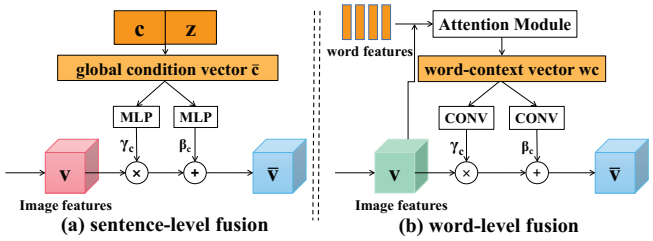

Figure 3: Semantic Fusion Module (SFM).

from excessively losing information, dilated convolution [Liu *et al.*, 2019] instead of traditional convolution is employed in a convolution block in each layer. In the semantic fusion module (SFM), semantic details at the sentence and word levels are integrated into the second and fourth, sixth and eighth layers of the encoder respectively. The decoder's structure mirrors that of the encoder, but the decoder doesn't have the SFM and uses the deconvolution operation. The generator undergoes training with text-matching loss, reconstruction loss, and inpainting loss, and edge loss explicitly. Furthermore, a one-stage generator is employed to reduce the parameters and the training time of the model.

3.2 Semantic Fusion Module

To make full use of the textual semantics, we propose a semantic fusion module (SFM), which aims to provide the semantic information in different granularities to the model and better guide this inpainting task. The SFM contains two components: sentence-level fusion and word-level fusion.

1) **Sentence-level Fusion:** The first four layers of the en-

coder are regarded as the coarse inpainting stage. Therefore, sentence-level fusion is applied in this stage. The condition vector c is derived from the sentence vector s using the condition enhancement method [Zhang *et al.*, 2017]. As shown in Figure 3(a), c and the noise $z \sim N(0,1)$ are concatenated to get a global condition $\overline{c}$. Then, the scale parameter γ_c and the shift parameter β_c are learned from the global condition $\overline{c}$ respectively, as shown in Eq.(1).

$$\gamma_c = W(\overline{c}), \beta_c = B(\overline{c}) \tag{1}$$

where $W(\cdot)$ and $B(\cdot)$ denote a linear projection layer respectively. The modulated visual features $\overline{v}$ is calculated by Eq.(2).

$$\overline{v} = \gamma_c \cdot \frac{v - \mu(v)}{\sigma(v)} + \beta_c \tag{2}$$

where v denotes the visual features of the image, $\mu(v)$ and $\sigma(v)$ denote the estimated average and variance, which are calculated from both batch and spatial dimensions for every channel.

2) **Word-level Fusion:** Naturally, the last four layers of the encoder are regarded as the fine inpainting stage. To generate more fine-grained inpainting results, word-level fusion is employed in this stage. w_i denotes the word embeddings of the i^{th} word. $v \in R^{C \times N}$ remains as the image-related hidden characteristics of the image, here, C represents the channel count and $N = W \times H$. The attention mechanism is used to compute a word-based context vector as the local condition for every sub-area of the image, as illustrated in Figure 3(b). Each column of v is a sub-area of the image. For the j^{th} sub-area, the word-based context wc_j is calculated by Eq.(3).

$$\mathrm{wc}_j = \sum_{i=0}^{T-1} \alpha_{j,i} w_i, \text{where } \alpha_{j,i} = \frac{\exp(h_{j,i})}{\sum_{k=0}^{T-1} \exp(h_{j,k})} \tag{3}$$

where $h_{j,i} = {v_j}^T w_i$, and $\alpha_{j,i}$ denotes the weight of i^{th} word w_i for the j^{th} sub-region v_j. After obtaining the local condition, two 1×1 convolution layers are adopted to convert wc into the word-level modulation parameters γ_c and β_c respectively. Finally, Eq.(2) is employed to modulate the visual hidden features v.

3.3 Word-Level Local Discriminator

To enhance the generator's ability to reconstruct missing patches based on textual descriptions, we introduce a word-level local discriminator, inspired by [Nam *et al.*, 2018; Li *et al.*, 2019]. Consistency is measured between the local regions and their corresponding word embeddings, rather than across the entire image.

As shown in Figure 4, there are two inputs to our word-level local discriminator: 1) the word embeddings $w \in R^{D \times T}$ obtained from the pre-trained RNN, and 2) v_{real} and v_{fake} are the local visual features corresponding to the real image I and generated image I' respectively. In the following, using $v \in R^{C \times N}$ to represent the local visual features

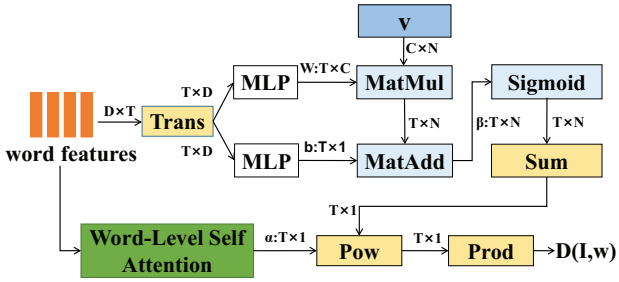

Figure 4: Details of Word-Level Local Discriminator (WLD). (Trans: Transpose; MLP: Linear Projection Layer; MatMul: Matrix Multiplication; Sum: Along Column Direction; Pow: element-wise; Prod: Along Row Direction.)

v_{real} and v_{fake}. Two linear projection layers are used to obtain the weight W and the bias b from the word embeddings w. For the j^{th} sub-region of the image v_j, Eq.(4) is applied to get a probability $\beta_{j,i}$ which can determine whether it is related to the i^{th} word w_i.

$$\beta_{j,i} = \sigma(W_i \cdot v_j + b) \tag{4}$$

where σ denotes the sigmoid function and $W_i \in R^{1 \times C}$. The self-attention at the word level [Nam *et al.*, 2018] is employed to compute the importance of the i^{th} word α_i relative to the temporal average of the word embeddings w. The final score $D(I,w)$ between the image I and word features w is calculated as bellow:

$$D(I,w) = \prod_{i=0}^{T-1} \left(\sum_{j=0}^{N-1} \beta_{j,i} \right)^{\alpha_i} \tag{5}$$

where N denotes the total count of sub-regions within the missing area.

3.4 Loss Functions

Let I, M_I, I' represent the original image, the damaged image, and the inpainting result respectively.

Reconstruction Loss: To guarantee that the inpainting result I' has a similar structure with the original image I, reconstruction loss is employed to measure the absolute error of each pixel, which is formulated as:

$$L_{rec} = \|I - I'\|_1 \tag{6}$$

Inpainting Loss: Inpainting Loss includes two parts: perceptual loss and total variation loss.

1) **Perceptual Loss:** Since reconstruction loss can not guarantee that the inpainting result has high-resolution textures, perceptual loss [Johnson *et al.*, 2016] is employed to obtain vivid textures. To extract high-level semantic features, a pre-trained VGG [Simonyan and Zisserman, 2015] network is utilized. The perceptual loss computes the absolute error of semantic features of the inpainting result I' compared to the original image I. The definition of perceptual loss is given as:

$$L_{per} = \sum_{j \in J} \|\phi_j(I) - \phi_j(I')\|_1 \tag{7}$$

where $\phi(\cdot)$ denotes the VGG network and J is selected VGG layers.

2) **Total Variation Loss:** Relying solely on perceptual loss results in checkerboard artifacts in the inpainting output [Johnson *et al.*, 2016], and [Johnson *et al.*, 2016] recommends using total variation loss to address this issue. Total variation loss L_{tv} calculates the error between each pixel in the inpainting image and its neighboring pixels above and to the left [Hong *et al.*, 2019].

The inpainting loss can be defined as:

$$L_{inp} = \lambda_{per} L_{per} + \lambda_{_{tv}} L_{tv} \tag{8}$$

where λ_{per} and λ_{tv} denote the weights of perceptual loss and total variation loss, respectively.

Edge Loss: To further generate vivid textures in the missing area, the edge loss is introduced. Edge information for the whole image is extracted using a Sobel operator. The errors are calculated from these edges instead of the original images. The edge loss is defined as:

$$L_{edge} = \|S(I) - S(I')\|_1 \tag{9}$$

where $S(\cdot)$ denotes a Sobel operator.

GAN Loss: To generate visually realistic images and missing patches which are consistent with words, the loss function for the generator is given by:

$$L_G = -E_{I' \sim \mathrm{p}_G}[\log(D(I')) + \log(D(I', w))] \tag{10}$$

The discriminator loss function is expressed as:

$$\begin{aligned} L_D = & -E_{I \sim \mathrm{p}_{data}}[\log(D(I)) + \log(D(I, w))] + \\ & - E_{I' \sim \mathrm{p}_G}[\log(1 - D(I')) + \log(1 - D(I', w))] \end{aligned} \tag{11}$$

where both $D(I, w)$ and $D(I', W)$ are computed by Eq.(5).

Finally, the loss function for CSF-GAN can be expressed as:

$$L_{total} = \lambda_{rec} L_{rec} + L_{inp} + \lambda_{\mathrm{egde}} L_{\mathrm{edge}} + \lambda_G L_G \tag{12}$$

where $\lambda_{rec}, \lambda_{inp}, \lambda_{edge}, \lambda_G$ serve as hyper-parameters to adjust the balance between the losses.

4 Experiments

We conduct a quantitative comparison between our CSF-GAN and the top-performing GAN-based inpainting techniques: RFR[Li *et al.*, 2020b], PD-GAN[Liu *et al.*, 2021] and PUT[Liu *et al.*, 2022], MMFL [Lin *et al.*, 2020], ALMR [Wu *et al.*, 2021], MIGT [Li *et al.*, 2023] and MISL[Wu *et al.*, 2024], which are recently proposed methods for text-guided image inpainting. Then, we carry out ablation studies to analyze the essential elements of our CSF-GAN, such as the Semantic Fusion Module (SFM), the Word-level Local Discriminator (WLD), and the edge loss.

4.1 Experiment Settings

Datasets: CSF-GAN is evaluated on two public benchmark datasets: CUB-200-2011 [Wah *et al.*, 2011] and Oxford-102 [Nilsback and Zisserman, 2008]. CUB-200-2011 includes 11788 bird images in 200 bird categories, and each image is paired with 10 corresponding textual descriptions. Following the settings of previous methods [Lin *et al.*, 2020; Wu *et al.*, 2021], we select 8,855 bird images from 150 species for training and 2,933 images from 50 species for testing. The Oxford-102 dataset comprises 102 flower species and 8,189 images in total, each of which has also 10 textual descriptions. We select 7034 and 1155 flower images with 82 categories and 20 categories as train and test sets respectively.

Methods	CUB-200-2011			Oxford-102		
	FID$^-$	PSNR$^+$	SSIM$^+$	FID$^-$	PSNR$^+$	SSIM$^+$
RFR[Li *et al.*, 2020b]	72.34	20.53	0.764	43.83	19.59	0.753
PD-GAN[Liu *et al.*, 2021]	65.94	18.27	0.688	49.27	17.27	0.651
PUT[Liu *et al.*, 2022]	56.19	18.33	0.727	42.58	16.95	0.678
MMFL[Lin *et al.*, 2020]	31.82	20.49	0.826	32.05	20.27	0.822
ALMR[Wu *et al.*, 2021]	32.33	16.23	0.507	32.65	20.32	0.812
MIGT[Li *et al.*, 2023]	35.30	**21.47**	**0.846**	32.48	**21.57**	**0.847**
MISL[Wu *et al.*, 2024]	14.77	18.87	0.764	30.85	0.764	0.726
CSF-GAN (ours)	**13.50**	20.75	0.829	**26.94**	20.61	0.834

Table 1: Performance comparing CSF-GAN with SOTA methods. $^-$ denotes better performance with lower values, and $^+$ with higher values.

Methods	Parameters (Million)	Training Time (Hour)
MMFL [Lin *et al.*, 2020]	149.20M	42.5 Hours
ALMR [Wu *et al.*, 2021]	175.87M	54.5 Hours
CSF-GAN (ours)	**82.29M**	**35.5 Hours**

Table 2: The number of trainable parameters and training time compared across different methods on the CUB-200-2011 dataset.

Implementation Details: CSF-GAN is implemented using PyTorch. All images in both bird and flower datasets are adjusted to a resolution of 256×256. In the image, a rectangular hole of size 128×128 is filled. A pre-trained bidirectional LSTM [Schuster and Paliwal, 1997] is used as the text encoder to extract sentence and word embeddings from the text. Each layer in the encoder contains a 3×3 convolution followed by a 4×4 dilated convolution. We use the Adam optimizer [Kingma and Ba, 2015] for parameter updates, setting $\beta_1 = 0.5, \beta_2 = 0.999$ and a learning rate of 0.0002. The hyper-parameters $\lambda_{rec}, \lambda_{per}, \lambda_{tv}, \lambda_{edge}$, and λ_G in the loss function are assigned values of 1.0, 0.5, 0.1, 0.5, 0.002 for both datasets respectively.

Evaluation Metrics: In line with prior studies [Lin *et al.*, 2020; Wu *et al.*, 2021], we assess the effectiveness of our CSF-GAN using several metrics: Fréchet Inception Distance (FID) [Heusel *et al.*, 2017]↓, Peak Signal-to-Noise Ratio (PSNR)↑, and Structural Similarity (SSIM) [Wang *et al.*, 2004]↑.

4.2 Comparison With the Baselines

Quantitative Results: To assess the effectiveness of CSF-GAN in text-guided image inpainting, we carry out numerous experiments comparing it with state-of-the-art inpainting techniques, such as RFR, PD-GAN, PUT, MMFL, ALMR, MIGT, and MISL. As shown in Table 1, CSF-GAN outperforms all compared methods in relation to FID on the CUB-200-2011 and Oxford-102 datasets, which can prove that inpainting images produced by our CSF-GAN exhibits superior image quality and greater visual plausibility. Compared to the best image inpainting method without text guidance, PUT, on the CUB dataset, CSF-GAN achieves a substantial

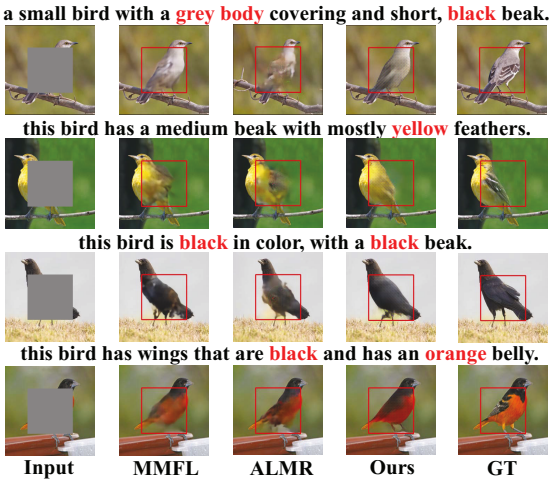

Figure 5: Visual examples comparing CSF-GAN with SOTA methods on the CUB-200-2011 dataset. The region within the red box is generated by the SOTA methods and CSF-GAN.

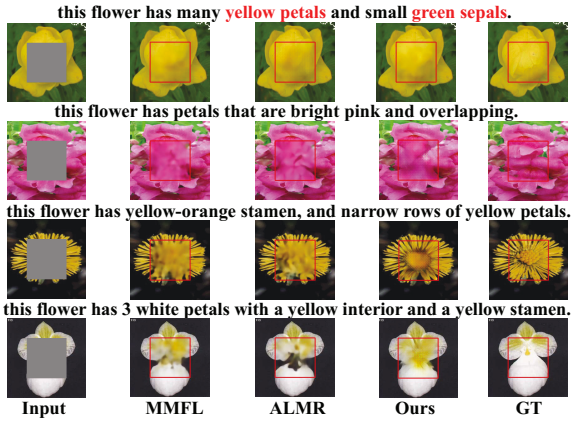

Figure 6: Qualitative examples of the SOTA methods and CSF-GAN on Oxford-102 flower dataset. The region within the red box is generated by the SOTA methods and CSF-GAN.

reduction in FID, from 56.19 to 13.50, achieving a remarkable improvement of 76.0%. On the Oxford-102 dataset, the FID score similarly declines from 42.58 to 26.94, marking an improvement of 36.73%. Although CSF-GAN shows slightly lower PSNR and SSIM scores than MIGT on both datasets, it still achieves competitive performance, ranking second overall. The reason lies in the fact that PSNR and SSIM evaluate the pixel-level variations between the original and inpainting results. We observe that high performance in PSNR and SSIM does not necessarily mean the ability to generate high-quality images [Zhao *et al.*, 2017]. For this reason, we do not include PSNR and SSIM results for these methods in subsequent assessments.

Furthermore, we conducted a comparison with two-stage text-driven image inpainting architectures, MMFL and ALMR, in terms of model parameters and training time. As shown in Table 2, although the parameters of these models are approximately twice as large as ours, and their training time is approximately 1.5 times longer than ours, our one-stage CSF-GAN model still demonstrated superior performance.

Qualitative Results: For evaluating the effectiveness of our CSF-GAN, we list some inpainting results generated by two SOTA text-guided image inpainting methods MMFL, ALMR, and CSF-GAN on two datasets in Figure 5 and Figure 6.

In the analysis of the image quality generated by the methods, the top row of Figure 5 and the fourth row of Figure 6 vividly demonstrate that the restoration results of MMFL and ALMR often appear blurred and structurally chaotic when the main object is obscured by a mask. In contrast, CSF-GAN not only accurately reconstructs the correct structure of bird and flower images but also produces visuals that are highly realistic and closely resemble the original images. Particularly, CSF-GAN displays a remarkable capability in detail reconstruction. For example, in the third line of Figure 5, guided by the text description "this flower has yellow-orange stamen," CSF-GAN endeavors to replicate the details of the stamens, manifesting the basic structure of the stamens clearly in its generated restoration result, while other models fail to identify specific objects in the damaged area clearly. This demonstrates CSF-GAN's strong ability to generate images rich in detail, surpassing existing text-guided image inpainting approaches.

Observing the restoration results in Figure 5 and Figure 6, it is evident that the images generated by CSF-GAN perfectly align with the text descriptions. While MMFL and ALMR can reflect the color information from the text descriptions in the restoration area to some extent, they lack depth in understanding the semantics of the text, particularly in accurately filling specific regions. For example, in the fourth row of Figure 5, the text description mentions "this bird has black wings and an orange belly." CSF-GAN more accurately interprets the semantics of the text, filling the matching textual information in the corresponding sub-areas of the image, resulting in a bird with black wings and an orange belly. This success is attributed to the efficient collaboration of our designed semantic fusion module and the word-level local discriminator.

Notably, the images inpainted by CSF-GAN may not show higher similarity to the original images, as the text descriptions may not include all the features. For example, the wings of the ground truth have black and white stripes in the first row of Figure 5, but the text does not mention this characteristic. Therefore, this may not lead to significant improvement in PSNR and SSIM.

4.3 Ablations

To assess the effectiveness of the proposed SFM and word-level local discriminator, and resolve their effects, we conduct ablation experiments on the CUB-200-2011 and Oxford-102 datasets. Due to our CSF-GAN being a one-stage method, we modify the MMFL as the baseline model. Table 3 summarizes the quantitative performance of the ablation study on both datasets. We discover that adding any component significantly boosts image quality. We incorporate the word-level local discriminator into the baseline model, which reduces the FID from 32.39 to 19.56 on the bird dataset and achieves 29.56 in FID on the flower dataset. This proves that the inclusion of both components enhances the performance of CSF-GAN.

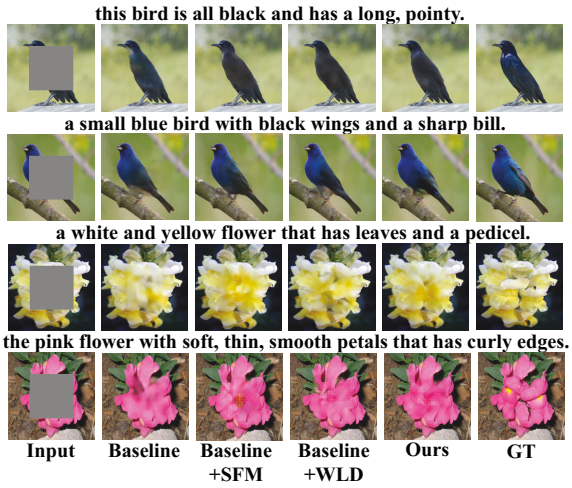

Figure 7: Visual comparisons of CSF-GAN variants on the CUB-200-2011 and flower datasets.

Methods	CUB-200-2011			Oxford-102		
	FID$^-$	PSNR$^+$	SSIM$^+$	FID$^-$	PSNR$^+$	SSIM$^+$
baseline	32.39	20.73	0.829	38.01	20.42	0.824
baseline+SFM	18.23	**20.89**	0.831	28.03	20.46	0.823
baseline+WLD	19.56	20.74	**0.832**	29.56	20.39	0.821
CSF-GAN(ours)	**13.50**	20.75	**0.829**	**26.94**	**20.61**	**0.834**

Table 3: Performance of different CSF-GAN components on the CUB-200-2011 and Oxford-102 datasets. $^-$ denotes better performance with lower values, and $^+$ with higher values.

As depicted in Figure 7, the baseline model generates blurred images. However, a significant change is observed when we introduce SFM or WLD into the baseline model. These enhanced models are capable of effectively parsing key information from text descriptions, thereby ensuring that the generated restoration results are closely aligned with the textual content. After adding SFM or WLD, the color of the generated birds becomes all black. This further proves that the introduction of SFM or WLD makes full use of the text semantics.

To prove the performance of the word-level fusion module, we visualize the attention map and show the top-4 most attended words by the attention map in Figure 8. The word-level fusion module and the WLD urge the model to focus more on important words (bright regions in Figure 8.). In Figure 8, we can observe that on the words related to attributes, such as colors, are emphasized for restoring details.

λ_{edge}	FID$^-$
no edge loss	30.23
0.1	15.87
0.2	19.23
0.5	**13.50**
1	16.31

Table 4: Influence of λ_{edge} on the efficiency of CSF-GAN on CUB-200-2011 dataset.

To demonstrate CSF-GAN's ability to repair images with irregular masks, Figure 9 presents several examples of inpainting results. The first row shows the irregular masks, and the second row displays the corresponding inpainted images. Notably, CSF-GAN generates realistic inpainting outcomes even for these challenging cases.

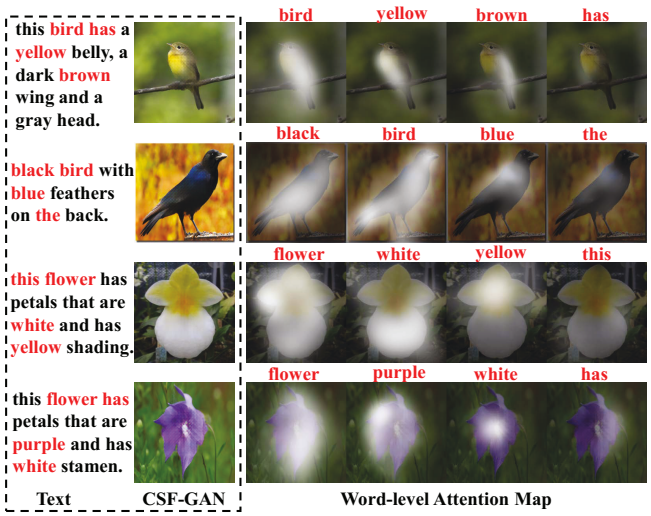

Figure 8: Examples of the word-level attention map on bird and flower datasets.

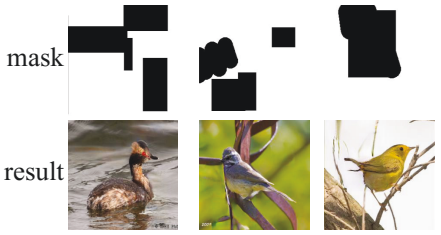

Figure 9: Results of repairing images with irregular masks

We also investigate how different values of λ_{edge} influence the performance of CSF-GAN. As shown in Table 4, it can be seen that changing the value of λ_{edge} has minimal impact on the performance of CSF-GAN. Finally, we set λ_{edge} to 0.5 as it has the best effectiveness on the bird dataset.

5 Conclusion

We present CSF-GAN, a novel one-stage text-guided image inpainting method. Our approach introduces two key innovations: (1) a semantic fusion module that integrates visual and textual features at varying granularities across different inpainting stages, and (2) a word-level local discriminator that ensures consistency between the generated inpainted patches and the corresponding words in the text. Additionally, we introduce an edge loss and an inpainting loss to regularize the inpainting process. Extensive experiments demonstrate that CSF-GAN not only achieves superior inpainting quality but also ensures strong alignment between the generated images and their associated textual descriptions.

Acknowledgments

This study is supported by Hainan Provincial Natural Science Foundation of China (No. 825CXTD608) and the grant (No. KYQD(ZR)-21079) and the Science and Technology Project of Liaoning Province (2024JH2/102600027) and the Science and Technology Project of Dalian City (2024JJ12GX025, 2023JJ12SN029 and 2023JJ11CG005).

References

[Chen *et al.*, 2022] Yizhou Chen, Xu-Hua Yang, Zihan Wei, Ali Asghar Heidari, Nenggan Zheng, Zhicheng Li, Huiling Chen, Haigen Hu, Qianwei Zhou, and Qiu Guan. Generative adversarial networks in medical image augmentation: a review. *Computers in Biology and Medicine*, 144:105382, 2022.

[Du *et al.*, 2024] Yingpeng Du, Di Luo, Rui Yan, Xiaopei Wang, Hongzhi Liu, Hengshu Zhu, Yang Song, and Jie Zhang. Enhancing job recommendation through llm-based generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8363–8371, 2024.

[Feng *et al.*, 2022] Xin Feng, Wenjie Pei, Fengjun Li, Fanglin Chen, David Zhang, and Guangming Lu. Generative memory-guided semantic reasoning model for image inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7432–7447, 2022.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017.

[Hong *et al.*, 2019] Xin Hong, Pengfei Xiong, Renhe Ji, and Haoqiang Fan. Deep fusion network for image completion. In Laurent Amsaleg, Benoit Huet, Martha A. Larson, Guillaume Gravier, Hayley Hung, Chong-Wah Ngo, and Wei Tsang Ooi, editors, *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 2033–2042. ACM, 2019.

[Iizuka *et al.*, 2017] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):107:1–107:14, 2017.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, 2016.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Li *et al.*, 2019] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Controllable text-to-image generation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 2063–2073, 2019.

[Li *et al.*, 2020a] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. Manigan: Text-guided image manipulation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7877–7886. Computer Vision Foundation / IEEE, 2020.

[Li *et al.*, 2020b] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7760–7768, 2020.

[Li *et al.*, 2023] Ailin Li, Lei Zhao, Zhiwen Zuo, Zhizhong Wang, Wei Xing, and Dongming Lu. MIGT: multimodal image inpainting guided with text. *Neurocomputing*, 520:376–385, 2023.

[Lin *et al.*, 2020] Qing Lin, Bo Yan, Jichun Li, and Weimin Tan. MMFL: multimodal fusion learning for text-guided image inpainting. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1094–1102. ACM, 2020.

[Liu *et al.*, 2019] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4169–4178. IEEE, 2019.

[Liu *et al.*, 2021] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9371–9381, 2021.

[Liu *et al.*, 2022] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. Reduce information loss in transformers for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11347–11357, 2022.

[Nam *et al.*, 2018] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial net-

works: Manipulating images with natural language. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 42–51, 2018.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society, 2008.

[Pathak *et al.*, 2016] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544. IEEE Computer Society, 2016.

[Reed *et al.*, 2016] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1060–1069. JMLR.org, 2016.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Wang *et al.*, 2004] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[Wu *et al.*, 2021] Xingcai Wu, Yucheng Xie, Jiaqi Zeng, Zhenguo Yang, Yi Yu, Qing Li, and Wenyin Liu. Adversarial learning with mask reconstruction for text-guided image inpainting. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran, editors, *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3464–3472. ACM, 2021.

[Wu *et al.*, 2024] Xingcai Wu, Kejun Zhao, Qianding Huang, Qi Wang, Zhenguo Yang, and Gefei Hao. MISL: multi-grained image-text semantic learning for text-guided image inpainting. *Pattern Recognition*, 145:109961, 2024.

[Xu *et al.*, 2018] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1316–1324. Computer Vision Foundation / IEEE Computer Society, 2018.

[Yin *et al.*, 2024] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, et al. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

[Zhang *et al.*, 2017] Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5908–5916. IEEE Computer Society, 2017.

[Zhang *et al.*, 2019] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1947–1962, 2019.

[Zhao *et al.*, 2017] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Image Processing*, 3(1):47–57, 2017.

[Zheng *et al.*, 2019] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 1438–1447. Computer Vision Foundation / IEEE, 2019.