

DAFT-GAN: Dual Affine Transformation Generative Adversarial Network for Text-Guided Image Inpainting

Jihoon Lee[†]

School of Electronic and Electrical Engineering
Kyungpook National University
Daegu, South Korea
leejh98123@knu.ac.kr

Yunhong Min[†]

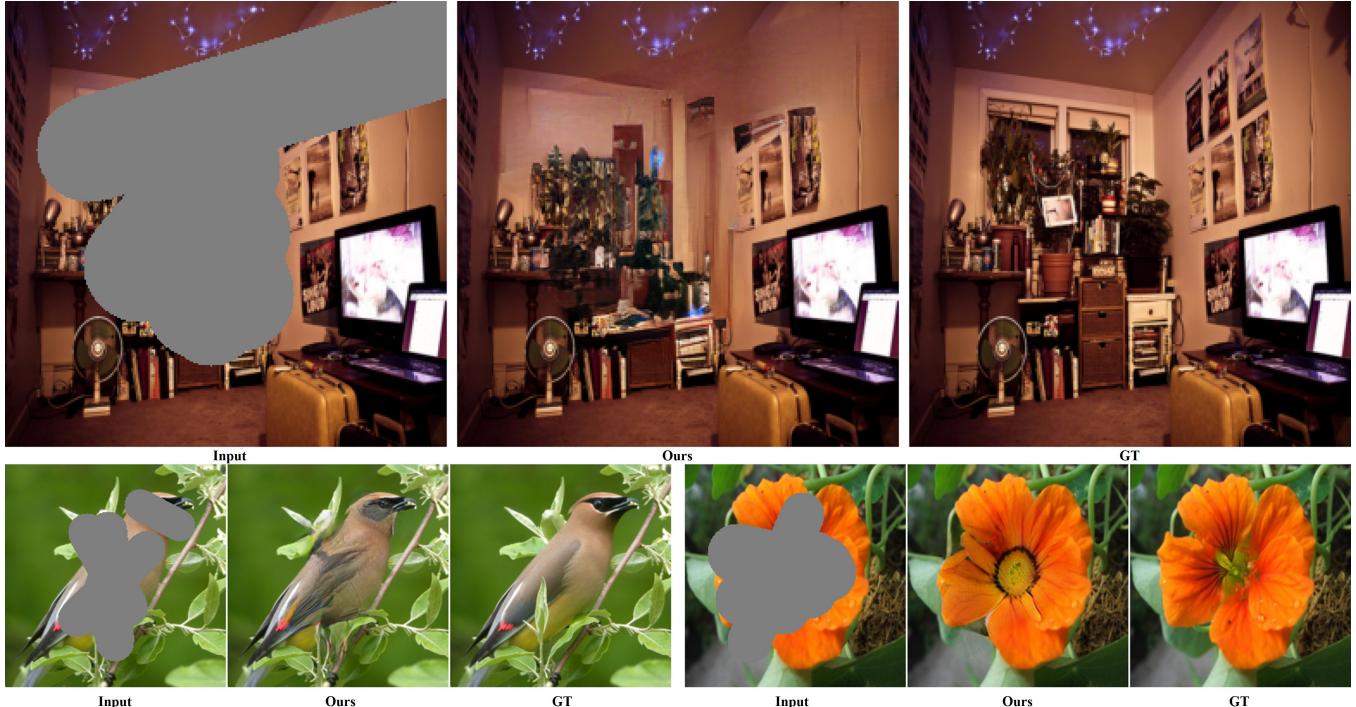
School of Electronics Engineering
Kyungpook National University
Daegu, South Korea
dbsghd363@knu.ac.kr

Hwidong Kim[†]

School of Electronics Engineering
Kyungpook National University
Daegu, South Korea
mat2408@knu.ac.kr

Sangtae Ahn^{*}

School of Electronic and Electrical Engineering
Kyungpook National University
Daegu, South Korea
stahn@knu.ac.kr



Text (top): "A room filled with furniture and accessories in a room.", **Text (bottom left):** "The bird is colorful and has black eyerings a spiky tan crown and gray wings.", **Text (bottom right):** "This is an orange flower with green stamen and black stripes near the ovary."

Figure 1: Results of proposed DAFT-GAN. Masked (left), generated (middle), and ground-truth (right) images are presented on three datasets (MS-COCO, CUB, and Oxford).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM'24, October 28 - November 1, 2024, Melbourne, Australia.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/18/06
<https://doi.org/XXXXXX.XXXXXXX>

ABSTRACT

In recent years, there has been a significant focus on research related to text-guided image inpainting. However, the task remains challenging due to several constraints, such as ensuring alignment between the image and the text, and maintaining consistency in distribution between corrupted and uncorrupted regions. In this paper, thus, we propose a dual affine transformation generative adversarial network (DAFT-GAN) to maintain the semantic consistency for text-guided inpainting. DAFT-GAN integrates two affine

transformation networks to combine text and image features gradually for each decoding block. Moreover, we minimize information leakage of uncorrupted features for fine-grained image generation by encoding corrupted and uncorrupted regions of the masked image separately. Our proposed model outperforms the existing GAN-based models in both qualitative and quantitative assessments with three benchmark datasets (MS-COCO, CUB, and Oxford) for text-guided image inpainting.

[†]These authors equally contribute to this work.

*Corresponding author: Sangtae Ahn (stahn@knu.ac.kr)

CCS CONCEPTS

- Computing methodologies → Artificial intelligence; Computer vision; Computer vision problems; Reconstruction.

KEYWORDS

Text-guided image inpainting, dual affine transformation, separated mask convolution, semantic consistency

ACM Reference Format:

Jihoon Lee[†], Yunhong Min[†], Hwidong Kim[†], and Sangtae Ahn*. 2024. DAFT-GAN: Dual Affine Transformation Generative Adversarial Network for Text-Guided Image Inpainting. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM'24), October 28–November 1, 2024, Melbourne, Australia*. ACM, New York, NY, USA, 9 pages. <https://doi.org/XXXXXXX-XXXXXX>

1 INTRODUCTION

Image inpainting [3, 8, 15–17, 22, 25, 32, 42] is a process that involves realistically filling in missing or corrupted areas within an image, serving a vital role in practical image processing applications such as image reconstruction, photo editing, and completing obscured regions. The core strategy in image inpainting involves predicting missing pixels by leveraging features or patterns from uncorrupted areas of the image. While these models can produce high-quality results, they face challenges in complex scenarios that demand intricate details, especially when multiple objects are involved or when the corrupted region covers a substantial portion of the image. Relying solely on surrounding pixels for generation in such cases can lead to distorted content or artifacts, prompting further research to address these limitations.

The pipeline typically follows a generative adversarial network (GAN) inpainting architecture, initially proposed in [20] and later refined by [5] to incorporate global and local aspects in the discriminator. However, addressing irregular mask holes has remained a challenge, prompting the introduction of a coarse-to-fine network [15] that includes a refinement process to enhance the generation quality. Additionally, previous studies [14, 34] proposed improved convolution techniques for feature encoding, while recent studies [26, 44] have demonstrated the benefits of utilizing fast Fourier convolutional (FFC) blocks to extract global and spatial features separately and modulate them individually for improved appearance quality. Nevertheless, accurately generating corrupted areas in complex images lacking specific patterns remains a significant challenge, as relying solely on surrounding pixels may yield semantically diverse solutions, necessitating time-consuming and

resource-intensive processes to select the desired output. In specific advanced inpainting tasks, therefore, incorporating external guidance to inpainting models is necessary to control the solution space and efficiently generate desired outcomes. Since external guidance, such as lines [14], edges [18], sketches [11], or exemplars [2, 6], may offer weak visual directionality and also lack semantic context, text guidance is considered the most effective form of guidance in inpainting tasks to ensure semantic consistency. Thus, text guidance is widely used in other tasks such as text-to-image synthesis and text-guided manipulation.

The task of text-guided inpainting is advancing towards generating high-quality images even in more complex cases such as filling in large masked regions in an image by utilizing text descriptions as external guidance. To do this, it is crucial to effectively utilize features from the uncorrupted regions of the image. Additionally, combining the features of the image with the text features used as guidance is also vital. The attention mechanism is the most commonly used when combining features from multimodal data. In text-to-image synthesis, the attention mechanism was first proposed in AttnGAN [31], demonstrating the effectiveness of combining text features with image features. Also, a deep attentional multimodal similarity model (DAMSM) [31] was proposed for extraction of sentence and word embeddings, which are crucial features when utilizing the attention mechanism. Similarly, in this study, we utilized the pre-trained DAMSM [31] to extract sentence and word embeddings. A recurrent affine network [33] was utilized when combining sentence embeddings with images, and a network with attention added to the existing affine network was employed for refinement at the word level when combining word embeddings with images. By gradually stacking this two-path decoder block, fine-grained images that effectively reflect text were generated. While most existing studies encode features of corrupted images through a single convolution, we propose separated mask convolution blocks to distinguish between corrupted and uncorrupted regions, minimizing the information leakage of uncorrupted image features. As shown in Fig. 1, Our proposed model successfully generates corrupted regions on three benchmark datasets.

The main contributions of this paper are listed as follows.

- We propose a novel text-guided inpainting model (DAFT-GAN) for generating fine-grained images with text descriptions.
- A dual affine transformation block is proposed to incorporate visual and text features effectively in the decoder stage from a global and spatial perspective.
- A separated mask convolution block is proposed to minimize the information leakage of uncorrupted image features.
- State-of-the-art performances are achieved on three benchmark datasets (MS-COCO, CUB-200-2011, and Oxford-102).

2 RELATED WORK

2.1 Text-to-Image Synthesis

Text-to-image synthesis is a task aimed at generating complete images using text descriptions. The most crucial aspects in text-to-image synthesis are the authenticity of the generated images and the semantic consistency between the provided text description and the generated images.

The foundation of text-to-image synthesis models, such as StackGAN [39], utilizes a multi-stage approach to generate high-resolution images reflecting the input text. This architecture not only gradually generates high-resolution images through stages but also proposes a stacked structure to address the unstable nature of GAN models. Subsequent research [31] focused on combining text and image features, with attention mechanisms being proven effective through such studies. However, attention mechanisms also have drawbacks when generating high-resolution images since they require attention at every scale, which incurs substantial computational costs. This computational burden can make it challenging to fully utilize text information at certain scales. To address these issues, DF-GAN [27] proposes an efficient fusion network that utilizes affine transformation networks without attention mechanisms. While this simple and efficient network can produce images with better quality, some limitations still exist. Some regions of the images may not be recognizable or consistent with the text description at the word level. To enhance this, SSA-GAN [12] introduces weakly supervised mask predictors to guide spatial transformations. RAT-GAN [33] adds recurrent networks when connecting affine transformation blocks to address the long-term dependency problem. Through research that effectively combines text and images, the quality of image generation has been improved.

2.2 Image Inpainting

Image inpainting involves the reconstruction of specific uncorrupted regions within an image, serving as a critical component in various image processing applications such as photo manipulation and filling in occluded areas. Despite its importance, image inpainting remains a challenging task, with the level of difficulty being heavily influenced by the extent of the damaged regions in the image. When a significant portion of the image is corrupted, there is a higher probability of entire objects being lost from the visual context. Moreover, the absence of usable features from uncorrupted areas further complicates the restoration process. Recent advancements in deep learning models have significantly enhanced the ability to extract and leverage high-level semantic features, resulting in notable improvements in terms of the quality of generated images. Recent studies investigated different model architectures, including transformer-based [9, 28, 35, 43], GAN-based [7, 26, 30, 36–38, 41, 44], or diffusion-based [10, 21, 23] models. Transformer and diffusion based models generally show good performance, but require a very high computational cost and time for training and inference than GAN-based models.

2.3 Text-Guided Image Inpainting

The goal of the text-guided image inpainting task is to reconstruct corrupted images using textual and visual information to generate realistic images. MMFL [13] proposes a multimodal fusion learning approach that focuses on text descriptions corresponding to objects of interest in each image through the word demand module. Additionally, they propose two-stage coarse-to-refine process to generate high-quality images. TDANet [40] proposes a dual multimodal attention module that aims to achieve deeper integration by applying attention to feature maps of two inverted regions. However, all these methods attempted multimodal fusion at the encoding stage and did not utilize the provided text information

during the decoding stage. This leads to an incomplete fusion of spatial details, text, and visual information. Furthermore, previous studies [7, 13, 40, 45] utilize structures with two or multiple stages for image refinement, resulting in significant time and space consumption due to repeated encoder-decoder pairs. Therefore, we designed a network that injects text description information at the decoding stage to generate images. This strategy not only mitigates resource inefficiencies but also improves image quality through the utilization of a one-stage-dual-path architecture.

3 METHODS

Given a masked image $I_M = I \odot M$ (where I is the original image, M is given mask metric), text-guided image inpainting is the task of generating an image that aligns with the text description t and maintains semantic consistency with the image. We propose DAFT-GAN, which generates images by appropriately processing text and visual features in the decoder stage.

3.1 Overall Architecture

As shown in Fig. 2, the proposed model consists of an encoder composed of separated mask convolution (SMC) blocks and a decoder composed of dual affine transformation (DAFT) blocks forming the Conv-U-Net structure. More specifically, the process of feature extraction involves seven SMC blocks, while text features are integrated at various resolutions through the incorporation of seven DAFT blocks to produce the final image. A comprehensive breakdown of all elements of the approach is provided as follows.

3.2 Separated Mask Convolution Block

Separated Mask Convolution. SMC block performs two roles. As shown in Fig. 3, one is to distinguish the masked and the unmasked regions and conduct convolution and normalization separately on them. Another is to update the mask in a way that minimizes information leakage. SMC block takes as input a pair of the image feature $F_i^e \in \mathbb{R}^{C_i \times W_i \times H_i}$ extracted from the previous block and a mask metric $M_i \in \{0, 1\}^{W_i \times H_i}$ of the same size. The mask metric represents the mask status, where the unmasked (valid) regions are marked as 0, and the masked (invalid) regions are marked as 1. The mask metric is used to differentiate between the two regions during the encoding process. The initial 256×256 masked image and mask metric are downscaled by a factor of 2 in 6 out of the 7 blocks, excluding the first block, resulting in a final output size of 4×4 .

First, from the previous block, we passed the input feature F_i^e through two different convolution layers to obtain a valid feature $F_{i+1}^{val} \in \mathbb{R}^{C_{i+1} \times \frac{W_i}{2} \times \frac{H_i}{2}}$ and an invalid feature $F_{i+1}^{inval} \in \mathbb{R}^{C_{i+1} \times \frac{W_i}{2} \times \frac{H_i}{2}}$. If we use only one convolutional layer to extract features, the convolutional weights could be updated by being influenced not only by the valid region but also by the invalid region during training. Therefore, we used two convolution layers.

$$F_{i+1}^{val} = \text{Norm}(\text{Conv}^{val}(F_i^e \odot (1 - M_i))), \quad (1)$$

$$F_{i+1}^{inval} = \text{Norm}(\text{Conv}^{inval}(F_i^e \odot M_i)). \quad (2)$$

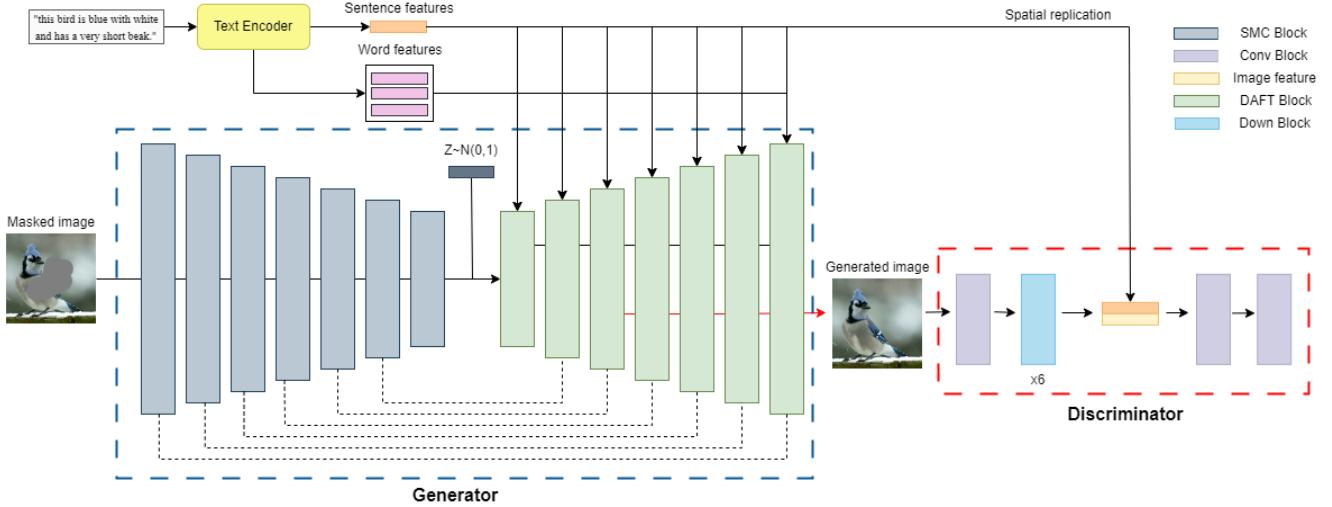


Figure 2: Architecture of DAFT-GAN consisting of an encoder-decoder generator and a one-way discriminator. The generator extracts image features and combines them with noise and text embeddings to generate the reconstructed images.

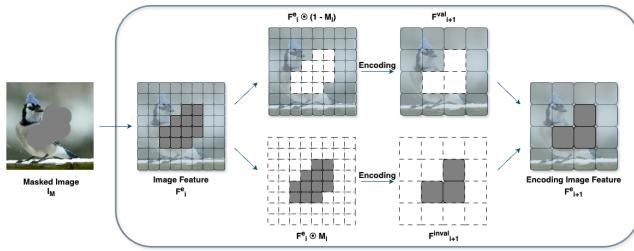


Figure 3: Visualization of the SMC block. Encoding includes convolution and normalization, generating higher-dimensional features that are downsampled by a factor of 2.

Then, we generated $F_{i+1}^e \in \mathbb{R}^{C_{i+1} \times \frac{W_i}{2} \times \frac{H_i}{2}}$ composed of F_{i+1}^{val} and $F_{i+1}^{invalid}$. At this point, using $M_{i+1} \in \{0, 1\}^{\frac{W_i}{2} \times \frac{H_i}{2}}$ obtained by updating M^i , we assigned F_{i+1}^{val} to the valid region of F_{i+1} and $F_{i+1}^{invalid}$ to the invalid region. The mask metric update process is explained in detail as follows.

$$F_{i+1}^e = F_{i+1}^{val} \odot (1 - M_{i+1}) + F_{i+1}^{invalid} \odot M_{i+1} \quad (3)$$

In F_{i+1}^e , the valid and invalid regions have distinctly different distributions. To prevent the loss of information in the valid region, we applied MaskNormalize, the normalization method used in [19].

Mask Update. When performing convolution operations on image features, both the valid and invalid regions are inevitably considered at the boundary of the mask due to the receptive field. In response to this situation, we applied a mask update method that treats as a valid feature if any part of the valid region is included to minimize the loss of valid information. We used a pooling layer with the same kernel size, stride, and padding as the convolution layer to extract the image features. By doing so, when performing convolution operations on a specific region of the image feature, we can also perform pooling operations on the exact corresponding region for the mask metric. If any part of the valid region is included in the receptive field, i.e., the region in the mask metric contains both 0

and 1, we need the pooling result for that region to be 0 in order to consider the convolution result as a valid feature. Therefore, we implemented a min-pooling operation to update the mask as follows.

$$M_{i+1} = -\text{MaxPool}(-M_i) \quad (4)$$

3.3 Dual Affine Transformation

Global Fusion Path. As depicted in Fig. 2, the generator generates masked parts by combining the encoded feature with the U-Net structure [24]. A noise vector z is sampled from a standard Gaussian distribution, passed through a fully connected layer, and then added to the encoded feature of the same dimension at the beginning of the generator. For each upsampling block, the encoded feature of the same scale was combined with the feature of the previous block through element-wise addition.

$$F_0^{gin} = MLP_1(z) \oplus F_L^e, \quad (5)$$

$$F_i^{gin} = F_{L-i}^e \oplus F_i^g, \quad (6)$$

where $1 \leq i \leq L$ (L is the highest level with the smallest spatial size)

Unlike the basic U-Net [24], this model combined features using residual connections instead of channel-wise concatenation. This was done to encourage the network output to closely resemble the input during the early stages of training, which stabilizes the learning process. Additionally, residual learning generally helps the model preserve previous information at each scale and learn high-frequency contents more effectively. The input F_i^{gin} passes through each recurrent affine transformation (RAT) module [33], performing the transformation for upsampling. It sequentially goes through the initial generator block and 6 upsampling blocks (the first generator block does not upsample), ultimately generating a feature map as a size of 256×256 .

Spatial Fusion Path. As shown in Fig. 4, multimodal cross affine transformation (MCAT) takes the RAT output image feature maps F_i^{gout} and word features w , hidden features h_t within the same DAFT block as input, and outputs same scale image feature maps

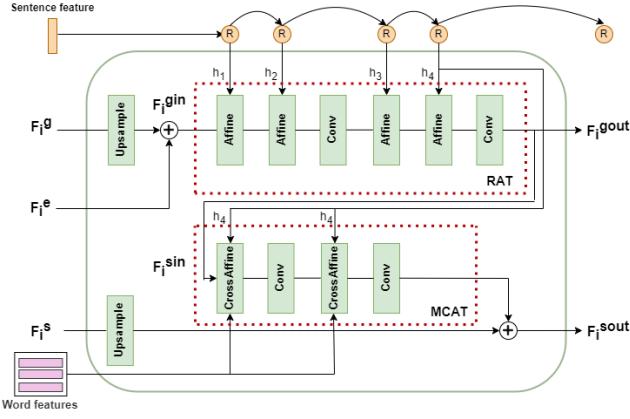


Figure 4: Structure of the DAFT block. The block is composed of RAT and MCAT modules, which respectively handle the global path and the spatial path, thereby forming a dual path architecture.

F_i^{sout} . The core of the MCAT module is the CrossAffine layer shown in Fig. 5. The CrossAffine layer performs cross-attention between image feature maps and word features, and obtains attention feature maps $F_{\text{spatial}} \in \mathbb{R}^{(2 \times d_w) \times H \times W}$ through channel-wise concatenation with hidden features h_t , and then extracts modulation parameters γ_c and β_c through channel-wise multi-layer perception (MLP) for affine transformation. By applying channel-wise MLP to predict modulation parameters while maintaining spatial structure and identifying where text information needs to be complemented in the current image features, spatially refined image feature maps can be obtained. Specifically, given input image feature maps $x_{chw} \in \mathbb{R}^{C \times H \times W}$, word features $w \in \mathbb{R}^{L \times d_w}$, and hidden feature $h_t \in \mathbb{R}^{d_w}$, the image feature maps were transformed into the same semantic space as word features by a new perceptron layer $W_Q \in \mathbb{R}^{d_w \times C}$. The query $Q_F = W_Q x_{chw} \in \mathbb{R}^{d_w \times H \times W}$, and key K_F , value V_F are word features w . Then, it performs cross-attention to obtain image feature maps F_{spatial} for predicting modulation parameters as follows:

$$F_{\text{attn}} = \text{softmax}(Q_F K_F^T) V_F, \quad (7)$$

$$F_{\text{spatial}} = [F_{\text{attn}}; \text{SpatialReplication}(h_t)], \quad (8)$$

where $\text{SpatialReplication}$ expands the dimensions by $H \times W$.

By combining the sentence hidden feature h_t from the long short-term memory with channel-wise concatenation with F_{attn} , necessary global semantic information at current stage was injected, at the same time, global-spatial connection was implemented. Additionally, modulation parameters γ_c and β_c were predicted through a channel-wise MLP, and channel-wise affine transformation and residual connection are applied to input feature maps x_{chw} to generate final output feature maps \tilde{x}_{chw} . This process can be represented by the following equation:

$$\text{CrossAffine}(x_{chw} | h_t, w) = \gamma_c x_{chw} + \beta_c, \quad (9)$$

where γ_c and β_c are obtained through two channel-wise MLP layers.

$$\gamma_c = \text{ChannelwiseMLP}_1(F_{\text{spatial}}), \quad (10)$$

$$\beta_c = \text{ChannelwiseMLP}_2(F_{\text{spatial}}), \quad (11)$$

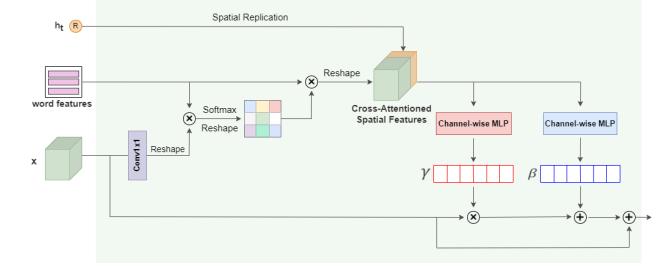


Figure 5: Diagram of CrossAffine module. The module manipulates input image features using recurrent hidden features and word features by channel wise affine transformation.

$$\tilde{x}_{chw} = x_{chw} \oplus \text{CrossAffine}(x_{chw} | h_t, w). \quad (12)$$

In order to learn a wider range of expressions, as described in Fig. 4, CrossAffine layer and convolution layer were stacked to form a single MCAT module.

One-Stage-Dual-Path. Inspired by recent research on image inpainting [44], we proposed a one-stage-dual-path approach that utilizes global text features to manipulate images, focusing on a global path for manipulating global structure and a spatial path for concentrating on spatial details. As shown in Fig. 4, the global path generates a semantic-consistent global structure through the RAT module, while the spatial path leverages more semantic-consistent and fine-grained visual details through word-level and pixel-level control using the MCAT module for each scale of the output from the global path. Additionally, by upsampling the previous spatial path output and connecting it to the current output through a residual connection, the model is guided to transform only the necessary parts gradually at each step, enabling initial learning stabilization and finer control over details.

This structure resolves the time and space waste issues derived from the two-stage or multi-stage structure of the existing text-guided image inpainting [7, 13, 45], and has the advantage of generating natural images from a human evaluation perspective through organic integration and detailed role differentiation between the two paths.

3.4 Objective Functions

3.4.1 Discriminator Objective.

Adversarial Loss with MA-GP. To ensure semantic consistency between the inferred image and the given text description, matching-aware zero-centered gradient penalty (MA-GP) was used [27].

$$\begin{aligned} L_D^{adv} &= E_{x \sim \mathbb{P}_{\text{data}}} [\max(0, 1 - D(x, s))] \\ &\quad + \frac{1}{2} E_{x \sim \mathbb{P}_G} [\max(0, 1 + D(\hat{x}, s))] \\ &\quad + \frac{1}{2} E_{x \sim \mathbb{P}_{\text{data}}} [\max(0, 1 + D(x, \hat{s}))] \\ &\quad + k E_{x \sim \mathbb{P}_{\text{data}}} [(\| \nabla_x D(x, s) \| + \| \nabla_s D(x, s) \|)^p], \end{aligned} \quad (13)$$

where s is the given text description, \hat{s} is the mismatched text description, x is the actual image corresponding to s , and \hat{x} is the generated image. $D(\cdot)$ is the output of the discriminator, providing

matching information between the image and the sentence. The k and p are hyperparameters of the MA-GP.

3.4.2 Generator Objective.

Reconstruction Loss. The ℓ_1 loss is typically optimized for an average blurry result. Despite being a non-saturating function, we incorporated perceptual loss to improve the naturalness and quality of the final image.

$$L_{rec} = \sum_i \sigma_i \|\phi_i(\hat{x}) - \phi_i(x)\|_2, \quad (14)$$

where σ_i is a weighting factor of each layer i and $\phi_i(\cdot)$ refers to the layer activation of the pre-trained VGG-19 network.

Adversarial Loss. Adversarial loss is defined as follows.

$$L_G = -E_{\hat{x}}[\log(D(x))], \quad (15)$$

where \hat{x} is generated images.

Text-Guided Attention Loss. To enhance text guidance, we implemented the text-guided attention loss [13]. This method involves multiplying the attention map and the generated image \hat{x} at the final scale of 256×256 with the ground-truth image x , and minimizing the ℓ_1 loss of the two terms.

$$L_{attn} = \|A(w, \hat{x})\hat{x} - A(w, \hat{x})x\|_1, \quad (16)$$

where $A(\cdot)$ performs attention and w is the word features of the text corresponding to x .

DAMSM Loss. For fine-grained image-text matching that considers both sentence-level and word-level information, we adopted the DAMSM loss (L_{DAMSM}). Details are described in [13].

Overall Loss. The total loss of the generator is defined as below.

$$L_G^{adv} = \lambda_{rec} \times L_{rec} + L_G + L_{attn} + \lambda_{DAMSM} \times L_{DAMSM} \quad (17)$$

4 EXPERIMENTS

4.1 Datasets

We used the Caltech-UCSD Birds-200-2011 (CUB-200-2011), Oxford-102 Category Flower (Oxford-102), and MS-COCO datasets to train our proposed model and methods, focusing on multimodal inpainting that incorporates both text and images. While all three datasets were used to assess the performance of text-guided inpainting, the MS-COCO dataset was specifically used as a more challenging dataset for evaluating the performance of image inpainting. This is because the MS-COCO dataset contains multiple objects and relatively complex scenes compared to the other two datasets.

4.2 Implementation Details

When training or testing our proposed model, we used randomly generated masks in irregular shapes. By doing so, we can construct a more robust model for inference and demonstrate the ability to generate realistic images even when large corrupted regions are present in the image. Subsequently, we evaluated the performance of the proposed model through several experiments. To ensure a fair evaluation, all models were retrained using the same corrupted masks. However, for the relatively recent GAN-based inpainting

Table 1: Performance comparison of inpainting models regarding FID, KID, PSNR, and SSIM. The evaluation was conducted on the test dataset in CUB-200-2011, Oxford-102, and MS-COCO.

Model	CUB-200-2011		Oxford-102		MS-COCO	
	FID↓	KID↓	FID↓	KID↓	FID↓	KID↓
MMFL [13]	25.68	1.266	35.69	1.560	19.77	0.803
TDA [40]	13.07	0.394	19.65	0.681	11.20	0.586
MIGT [7]	35.30	-	32.48	-	-	-
MISL [30]	14.77	0.296	30.85	0.763	23.34	1.273
SD-Inpainting [23]	27.46	2.204	32.60	2.116	7.48	0.406
SD-Inpainting-XL [21]	14.08	0.633	28.74	1.970	7.13	0.386
GLIGEN [10]	12.01	0.266	17.42	0.487	6.97	0.364
Ours	11.33	0.259	15.75	0.318	6.59	0.357
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MMFL [13]	20.34	0.799	20.61	0.811	20.48	0.769
TDA [40]	20.23	0.797	19.02	0.753	19.56	0.652
MIGT [7]	21.47	0.846	21.57	0.847	-	-
MISL [30]	18.87	0.764	17.47	0.726	18.11	0.734
SD-Inpainting [23]	18.46	0.708	17.92	0.680	18.12	0.685
SD-Inpainting-XL [21]	23.49	0.826	21.76	0.761	22.76	0.798
GLIGEN [10]	18.21	0.720	18.37	0.715	18.63	0.741
Ours	20.46	0.808	20.25	0.766	19.64	0.783

Table 2: Numerical ranking scores for semantic consistency and naturalness. Semantic consistency and naturalness indicate the alignment between text and image, and the quality of the image, respectively.

Model	Semantic consistency	Naturalness	Param(B)	Time(sec/image)
MMFL [13]	3.392	4.650	0.15	0.317
TDA [40]	2.287	3.089	0.03	0.192
SD-Inpainting [23]	2.049	2.769	1.07	8.62
SD-Inpainting-XL [21]	1.505	1.643	2.77	18.9
GLIGEN [10]	1.725	2.111	1.07	17.49
Ours	1.935	2.357	0.09	0.029
Ground-truth	1.19	1.112		

models, such as MIGT [7] and MISL [30], we utilized the quantitative results reported in the previous papers(MIGT and MISL), as the official source code was not publicly available. We utilized training approach with diverse irregular masks. We conducted the experiment on an NVIDIA RTX3090 Ti GPU. The parameters of the generator in the model were optimized with a learning rate of 10^{-4} , while the parameters of the discriminator were optimized with a learning rate of 4×10^{-4} , both using the Adam optimizer. The sentence embedding and word embedding utilized in this study were extracted using a pre-trained text encoder from the DAMSM proposed in AttnGAN [31]. The final objective function weights were set as $\lambda_{rec} = 0.2$ and $\lambda_{DAMSM} = 0.01$

4.3 Quantitative Results

The image inpainting task allows for evaluating the quality of generated images by comparing how well the corrupted parts are reconstructed to the ground truth. To assess the quality of generated images, the Fréchet inception distance (FID) [4] and the kernel inception distance (KID) [1] were used to measure the distribution between ground-truth and generated images. When evaluating reconstruction, the peak signal-to-noise ratio (PSNR) and the structural similarity index (SSIM) [29] were used to measure the difference between generated images and ground-truth at the pixel level. Lower values for FID and KID, and higher values for PSNR and SSIM indicate better performance. As shown in Table 1, we observed



Figure 6: Qualitative comparison of ours with other inpainting models on the CUB-200-2011 dataset.

Table 3: Performance comparison of each component added. To compare three different affine transformation methods, the baseline (DF-GAN) was compared against methods that applied the SSA and RAT module.

Methods	FID↓	KID↓	PSNR↑	SSIM↑
Our Baseline (DF-GAN)	19.31	0.759	18.15	0.790
Our Baseline + SSA	18.79	0.730	19.80	0.796
Our Baseline + RAT	17.25	0.593	19.87	0.806
Our Baseline + MCAT	17.84	0.641	20.13	0.814
Our Baseline + DAFT (MCAT + RAT)	14.23	0.418	18.95	0.757
Our Baseline + DAFT + SMC	12.57	0.293	19.06	0.742
Our Final (Diverse Mask + SMC + DAFT)	11.33	0.259	20.46	0.808

that our proposed model has a significant impact on integrating text and achieving high scores compared to the existing models on all three datasets, with a particularly significant improvement seen on the MS-COCO dataset. Despite being relatively challenging due to its inclusion of multiple objects and complex scenes, the significant improvement on the MS-COCO dataset indicates effective semantic matching between text and images. Despite the significant improvements in FID and KID scores across all datasets, the improvements in PSNR and SSIM are relatively modest. This phenomenon may be explained by the distinction that PSNR and SSIM metrics evaluate distances at the pixel level, while our model operates on feature-level distances to produce images that exhibit a more realistic appearance. Consequently, the improvement in reconstruction metrics compared to previous models that rely solely on pixel-level distances might be lower. Nonetheless, it is evident that in some cases, other models achieve similar or even higher reconstruction scores.

4.4 Qualitative Results

In the context of generative models, qualitative evaluation from a human perspective is also important. In the text-guided image inpainting task, the most commonly used metrics for qualitative evaluation are semantic consistency and naturalness. Semantic consistency measures how well the generated image matches the given

text, while naturalness measures how natural the generated image appears. We conducted a human evaluation with 15 volunteers with 100 randomly generated images, ranging from rank 1 (high quality) to rank 5 (low quality), to obtain scores for semantic consistency and naturalness. As scored in Table 2, we observed that our model outperforms not only GAN-based inpainting models [13, 40] but also diffusion-based models, such as Stable Diffusion Inpainting model [23], in both quantitative and qualitative evaluations. As evident from the metrics in Fig. 6, while other GAN-based models generated awkward images with blurriness or artifacts, our model generated natural-looking images without any anomalies compared to the ground truth. To ensure high-quality image generation across all datasets, we conducted experiments with various text descriptions using our proposed model. In Fig. 7, we confirmed the generation of photorealistic images on the CUB and the Oxford datasets. Particularly in Fig. 8, despite the diverse object types and more complex scenes, our model effectively incorporated the text descriptions to generate remarkably natural images, even for the challenging MS-COCO dataset.

4.5 Comparisons with Diffusion-based Models

Recent diffusion-based models, such as Stable Diffusion and GLIGEN, have shown significant advancements in image quality in the context of image generation. However, diffusion models, trained on large datasets, learn general features, which may not be ideal for evaluations on specific datasets such as CUB-200-2011 and Oxford-102. As shown in Fig. 6 and Table 1, our proposed GAN-based model outperforms diffusion models (SD-Inpainting [23], GLIGEN [10], and SD-Inpainting-XL [21]) in minimizing distribution differences between undamaged and damaged parts of an image, resulting in the generation of more realistic images. However, while our model achieved better quantitative metrics compared to the most recent diffusion-based models, it scored relatively lower in qualitative assessments, particularly in human evaluation. Although it might be considered unfair to compare our model with diffusion-based models trained on large-scale datasets, rather than GAN-based approaches, such comparisons are essential to demonstrate the effectiveness and competitiveness of our inpainting model. It is important to note that, while qualitative evaluation is inherently more subjective than quantitative evaluation, the fact that our model falls short in terms of overall performance, including image quality, compared to diffusion-based models with massive parameters and datasets, represents a limitation but our proposed model is advantageous in terms of the number of parameters and the efficiency (inference time) of the inpainting process as shown in Table 2.

4.6 Effectiveness of Each Proposed Method

In Table 3, we conducted an ablation study on the CUB-200-2011 dataset to assess the performance changes resulting from sequentially applying our proposed methods to the baseline model. The baseline model utilizes a basic convolutional block to encode the features of the masked image and employs only one affine transformation network [27] along with sentence features. From this baseline, we measured the performance changes when adding the weakly supervised mask predictor proposed in [12], and when adding the recurrent affine network proposed in [33]. Then, we

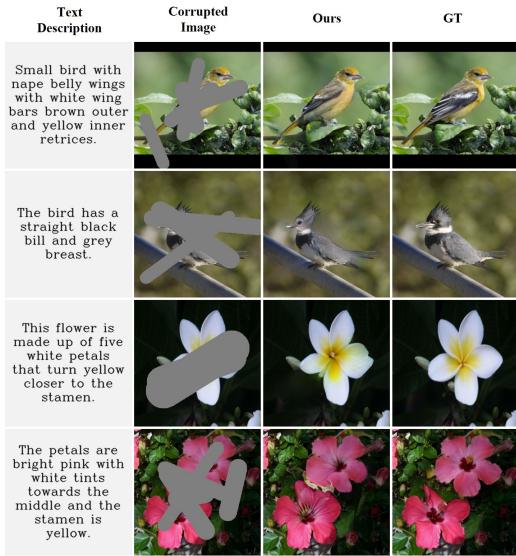


Figure 7: Results of our proposed model on CUB-200-2011 (first and second rows) and Oxford-102 (third and fourth rows) datasets. Corrupted (left), generated (middle), and ground-truth (right) images are presented.

evaluated the performance when adding cross refinement, which involves using two affine transformation networks to refine by integrating word features. Additionally, we evaluated the performance by incorporating the DAFT Block, which connects the features of each encoder to the decoder using residual connections. Then, we examined the impact of replacing the basic block with the SMC block in our final model. Finally, to build a robust inpainting model that is independent of the shape of corrupted masks, we trained our model using various irregular masks. By sequentially applying our proposed methods, we confirmed significant improvements in performance based on the evaluation metrics.

4.7 Text Controllability for Image Manipulation

Through text guidance, we can manipulate specific parts to create desired images or generate diverse images that are semantically consistent with various text descriptions. We confirmed it through the following process. First, mask the part of the image we want to manipulate. Second, provide a textual description of the anticipated image post-manipulation. Third, feed the masked image and text into the model. As a result, we presented two different results when two different sentences were inserted into each image. As shown in Fig. 9, the first and second examples utilize different images and masks to show color and size variations, as well as the inclusion of relevant semantic information in generated images. The results show the ability to accurately locate and generate modifications based on the provided text description.

5 CONCLUSIONS

In this study, we propose DAFT-GAN to address the current challenges for text-guided image inpainting. Our approach involves the integration of two affine transformation networks to progressively

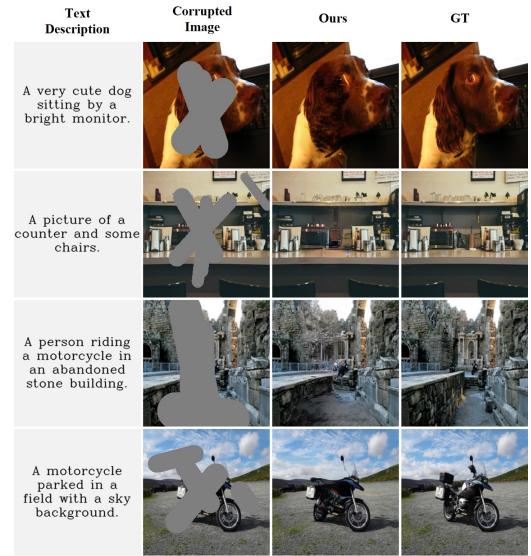
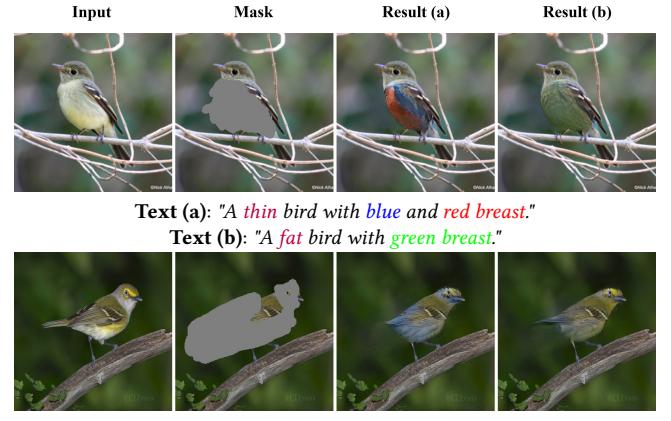


Figure 8: Results of our proposed model on MS-COCO Dataset. Corrupted (left), generated (middle), and ground-truth (right) images are presented.



Text (a): "A **thin** bird with **blue** and **red** **breast**."
Text (b): "A **fat** bird with **green** **breast**."

Text (a): "A **tiny** bird with **blue** **heads** and **short** **blue** **tail**."
Text (b): "A **tiny** bird with **yellow** **heads** and **long** **black** **tail**."

Figure 9: Generated images with different text guidance.

incorporate text and image features, thereby enhancing the semantic consistency between the generated images and associated text descriptions. Using global text features, our model initially generates coarse results, which are subsequently refined using spatial details, leading to an overall quality improvement of the generated images. Additionally, we propose SMC blocks to reduce information leakage of uncorrupted features by encoding corrupted and uncorrupted regions of the masked image separately. Through experiments, our method outperforms existing GAN-based methods in both qualitative and quantitative evaluations which represent significant advancements in text-guided image inpainting.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022R1A4A1023248 and No. RS-2023-00209794).

REFERENCES

- [1] Mikolaj Biorkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying mdm gans. *arXiv preprint arXiv:1801.01401* (2018).
- [2] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing* 13, 9 (2004), 1200–1212.
- [3] Soheil Darabi, Eli Shechtman, Connally Barnes, Dan B Goldman, and Pradeep Sen. 2012. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on graphics (TOG)* 31, 4 (2012), 1–10.
- [4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).
- [5] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1–14.
- [6] Vivek Kwatra, Irfan Essa, Aaron Bobick, and Nipun Kwatra. 2005. Texture optimization for example-based synthesis. In *ACM SIGGRAPH 2005 Papers*. 795–802.
- [7] Ailin Li, Lei Zhao, Zhiwen Zuo, Zhizhong Wang, Wei Xing, and Dongming Lu. 2023. MIGT: Multi-modal image inpainting guided with text. *Neurocomputing* 520 (2023), 376–385.
- [8] Jingyu Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. 2020. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7760–7768.
- [9] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10758–10768.
- [10] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. *CVPR* (2023).
- [11] Jiamin Liang, Xin Yang, Yuhao Huang, Haoming Li, Shuangchi He, Xindi Hu, Zejian Chen, Wufeng Xue, Jun Cheng, and Dong Ni. 2022. Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. *Medical image analysis* 79 (2022), 102461.
- [12] Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. 2022. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18187–18196.
- [13] Qing Lin, Bo Yan, Jichun Li, and Weimin Tan. 2020. Mmfl: Multimodal fusion learning for text-guided image inpainting. In *Proceedings of the 28th ACM international conference on multimedia*. 1094–1102.
- [14] Guolin Liu, Fitzsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*. 85–100.
- [15] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. 2019. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4170–4179.
- [16] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. 2021. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9371–9381.
- [17] Yuqing Ma, Xianglong Liu, Shihao Bai, Lei Wang, Dailan He, and Aishan Liu. 2019. Coarse-to-Fine Image Inpainting via Region-wise Convolutions and Non-Local Correlation. In *Ijcai*. 3123–3129.
- [18] K Nazeri, E Ng, T Joseph, FZ Qureshi, and M Ebrahimi. 2020. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* 2019. *arXiv preprint arXiv:1901.00212* (2020).
- [19] Minheng Ni, Xiaoming Li, and Wangmeng Zuo. 2023. NUWA-LIP: language-guided image inpainting with defect-free VQGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14183–14192.
- [20] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2536–2544.
- [21] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
- [22] Chuan Qin, Shuzhong Wang, and Xinpeng Zhang. 2012. Simultaneous inpainting for image structure and texture using anisotropic heat transfer model. *Multimedia Tools and Applications* 56 (2012), 469–483.
- [23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.
- [24] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18. Springer, 234–241.
- [25] Yuhang Song, Chaoyang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. 2018. Contextual-based image inpainting: Infer, match, and translate. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.
- [26] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2149–2159.
- [27] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. 2022. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16515–16525.
- [28] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. 2021. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4692–4701.
- [29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
- [30] Xingcai Wu, Kejun Zhao, Qianding Huang, Qi Wang, Zhenguo Yang, and Gefei Hao. 2024. MISL: Multi-grained image-text semantic learning for text-guided image inpainting. *Pattern Recognition* 145 (2024), 109961.
- [31] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.
- [32] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. 2018. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*. 1–17.
- [33] Senmao Ye, Huan Wang, Mingkui Tan, and Fei Liu. 2023. Recurrent affine transformation for text-to-image synthesis. *IEEE Transactions on Multimedia* (2023).
- [34] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4471–4480.
- [35] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. 2021. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*. 69–78.
- [36] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. 2022. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics* (2022).
- [37] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M Patel. 2021. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 14164–14173.
- [38] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. 2020. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX* 16. Springer, 1–17.
- [39] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.
- [40] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. 2020. Text-guided neural image inpainting. In *Proceedings of the 28th ACM international conference on multimedia*. 1302–1310.
- [41] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. 2021. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428* (2021).
- [42] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1438–1447.
- [43] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2021. Tfll: Image completion via a transformer-based architecture. *arXiv preprint arXiv:2104.00845* 2, 3 (2021), 6.
- [44] Haitian Zheng, Zhe Lin, Jingwan Lu, Scott Cohen, Eli Shechtman, Connally Barnes, Jianming Zhang, Ning Xu, Sohrab Amirghodsni, and Jiebo Luo. 2022. Image inpainting with cascaded modulation gan and object-aware training. In *European Conference on Computer Vision*. Springer, 277–296.
- [45] Yucheng Zhou and Guodong Long. 2023. Improving cross-modal alignment for text-guided image inpainting. *arXiv preprint arXiv:2301.11362* (2023).

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58

Supplementary Materials: Diverse Generation

Anonymous Authors

1 DIVERSE IMAGE GENERATION

We propose DAFT-GAN, an effective approach for integrating text and images. To validate the performance of our proposed model, we utilized the CUB-200-2011, Oxford-102, and MS-COCO datasets. Through both quantitative and qualitative evaluations, we verified that our proposed model outperforms other models. Furthermore, when conducting inference on a variety of classes and complex scenes, not limited to simple scenes or specific classes, we visually confirmed that our model largely produces natural and realistic images.

As shown in Fig.1, inference was performed on the CUB dataset, in Fig.2 on the Oxford-102 dataset, and in Fig.3 on the COCO dataset. While our model excels at generating natural images for CUB and Oxford-102, it particularly stands out in generating high-quality images across diverse scenarios, including challenging classes and complex scenes, within the COCO dataset. Previous inpainting models often struggled to produce high-quality results in challenging scenarios or complex scenes, such as generating images of people or intricate patterns. However, our model demonstrates good quality generation across diverse situations, including animals, humans, and complex indoor environments.

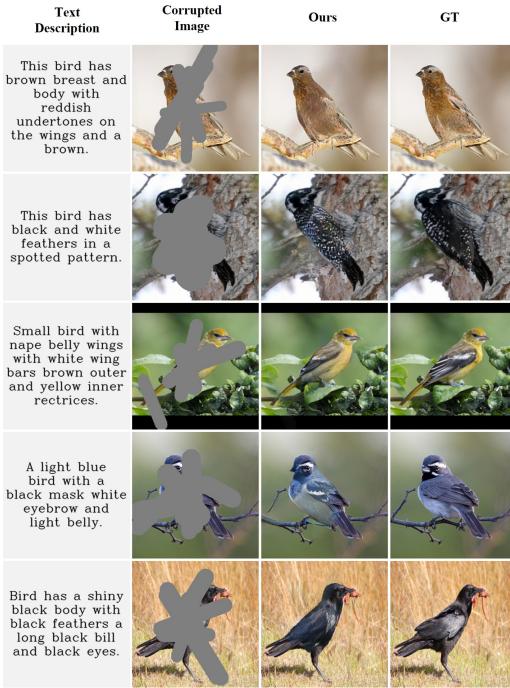


Figure 1: Results of our proposed model on CUB-200-2011 dataset. Corrupted (left), generated (middle), and ground-truth (right) images are presented.

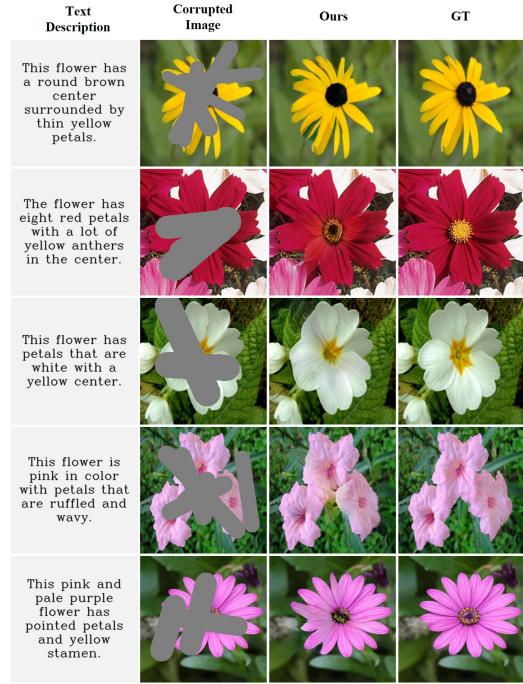


Figure 2: Results of our proposed model on Oxford-102 dataset. Corrupted (left), generated (middle), and ground-truth (right) images are presented.

2 IMPORTANCE OF TRAINING MASK DIVERSITY FOR ENHANCING ROBUSTNESS

When conducting quantitative and qualitative evaluations of our proposed DAFT-GAN as well as other inpainting models, we trained all comparison models using diverse irregular masks with a wide range of masking ratios from around 10% to 70% to improve its robustness. This is also aimed at solely focuses on evaluating performance and effectiveness based on the inpainting method. In contrast, using only biased masks can make the model's robustness highly vulnerable, as shown in Fig.4. The previous MMFL trained the model using only 25% center masks, and the results were satisfactory when evaluated on the same 25% center masks, as shown in Fig.5. However, when the same model was evaluated with diverse irregular masks, the overall results became poor, and the areas outside the center mask that did not need to be recovered during training became completely degraded, essentially becoming noise-like. This demonstrates that when training with biased masks, the model's weights also become biased towards only being able to recover for those specific mask patterns. For the experiments in this paper, we directly trained and evaluated the other models, including MMFL, using the same diverse irregular masks, to enable an accurate comparative analysis.

59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116

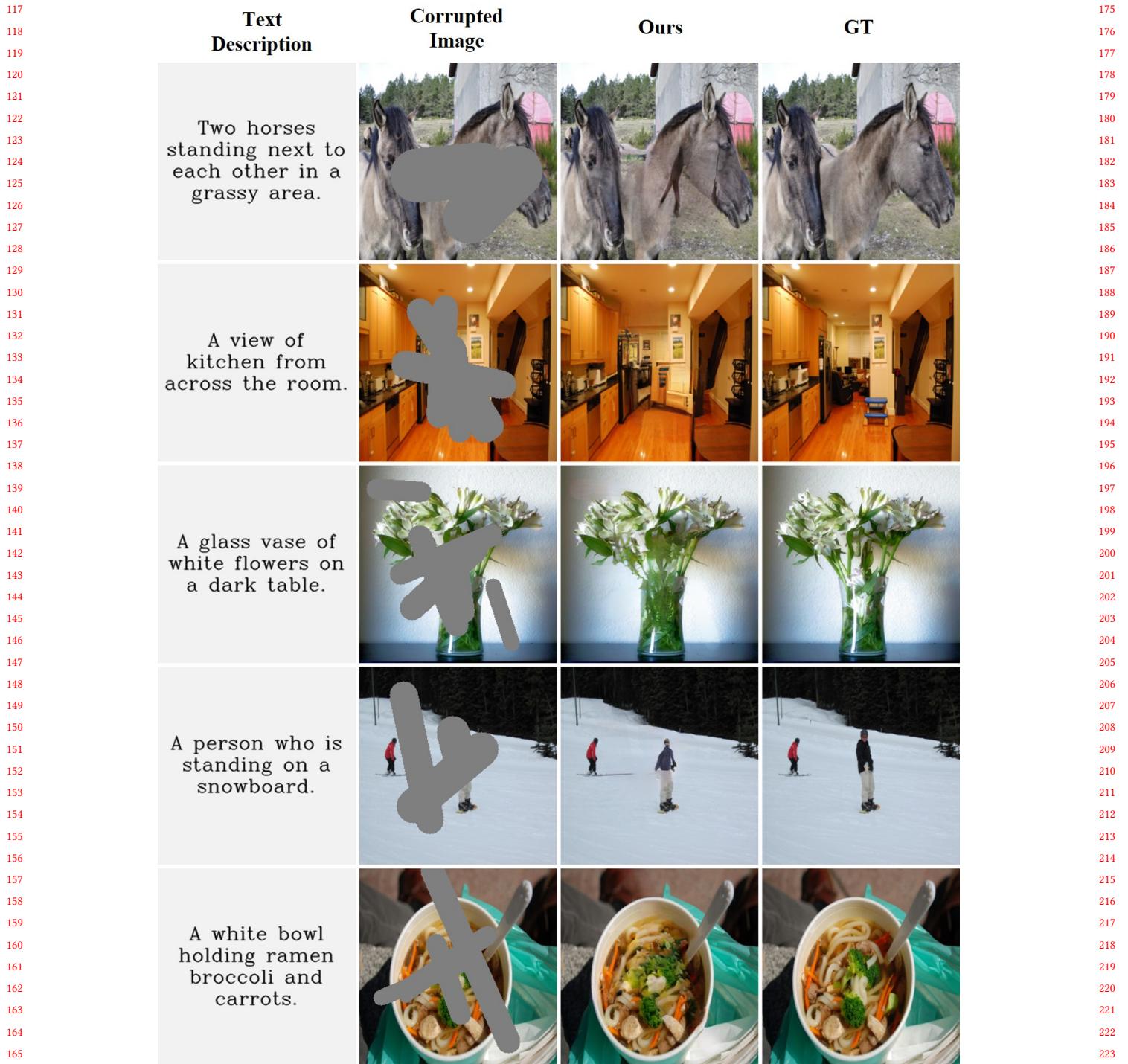
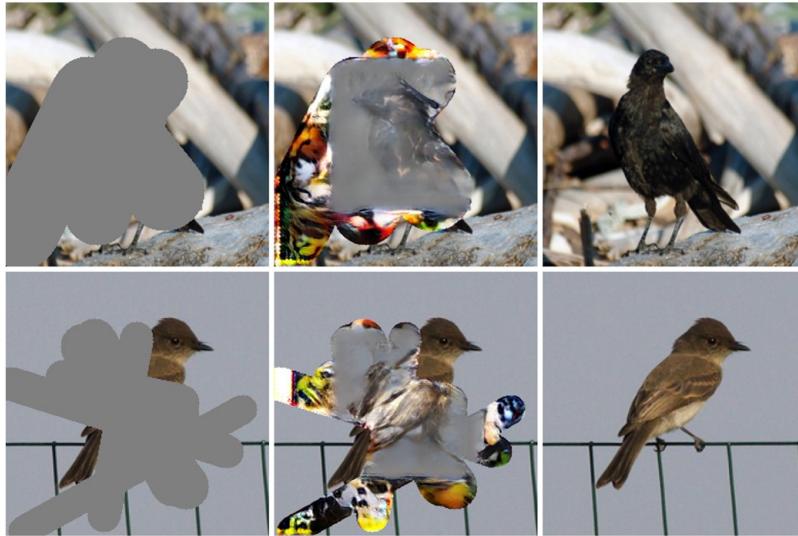
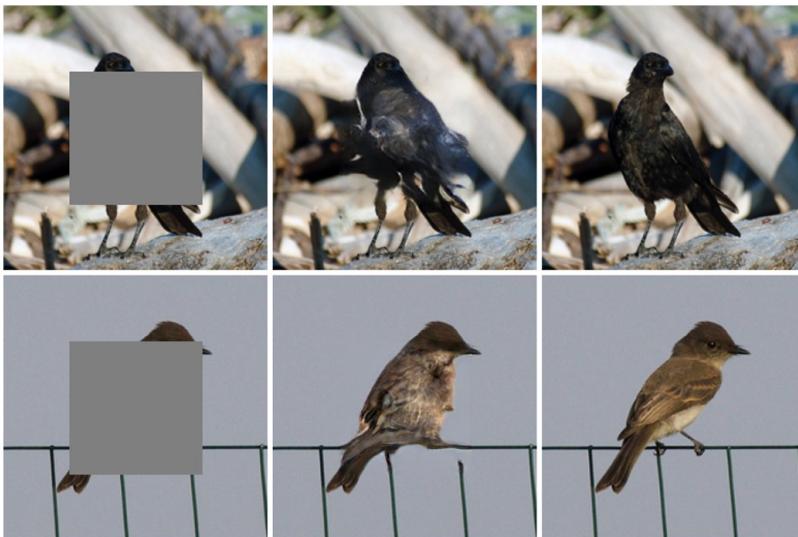


Figure 3: Results of our proposed model on MS-COCO dataset. Corrupted (left), generated (middle), and ground-truth (right) images are presented.

233
234
235**Corrupted
Image****Generated****GT**291
292
293**Text (a):** "This bird has wings that are black and has a thick bill."**Text (b):** "A small bird with dull feathers and a little mohawk on top of head."255
256
257**Figure 4:** Generated images with diverse irregular masks, trained with center masks. Corrupted (left), generated (middle), and ground-truth (right) images are presented. Text(a) (first row), Text(b) (second row) are demonstrated.313
314
315

260

**Corrupted
Image****Generated****GT**316
317
318**Text (a):** "This bird has wings that are black and has a thick bill."**Text (b):** "A small bird with dull feathers and a little mohawk on top of head."283
284
285**Figure 5:** Generated images with center masks, trained with center masks. Corrupted (left), generated (middle), and ground-truth (right) images are presented. Text(a) (first row), Text(b) (second row) are demonstrated.341
342
343288
289
290