Text-Guided Neural Image Inpainting

Lisai Zhang[1], Qingcai Chen*[1,2], Baotian Hu[1], Shuoran Jiang[1]
[1]Shenzhen Chinese Calligraphy Digital Simulation Engineering Laboratory, Harbin Institute of Technology, Shenzhen
[2]Peng Cheng Laboratory, Shenzhen, China
LisaiZhang@foxmail.com, qingcai.chen@hit.edu.cn*, {baotianchina,shaunbysn}@gmail.com

Origin | Corrupted | Without Text | **Text-guided (a)** | **Text-guided (b)** | **Text-guided (c)**

Text (a): *"This bird has gray head, black and white wings and yellow belly."* **Text (b)**: *"This bird is gray in color, with red belly."*
Text (c): *"A small bird with spotted blue wings, gray tail and head, and has white throat, breast, belly and white undertail."*

Figure 1: Illustration of inpainting a unique area. (a) is guided inpainting case with the original image description, (b) and (c) produces different new contents while guided with altered texts.

ABSTRACT

Image inpainting task requires filling the corrupted image with contents coherent with the context. This research field has achieved promising progress by using neural image inpainting methods. Nevertheless, there is still a critical challenge in guessing the missed content with only the context pixels. The goal of this paper is to fill the semantic information in corrupted images according to the provided descriptive text. Unique from existing text-guided image generation works, the inpainting models are required to compare the semantic content of the given text and the remaining part of the image, then find out the semantic content that should be filled for missing part. To fulfill such a task, we propose a novel inpainting model named Text-Guided Dual Attention Inpainting Network (TDANet). Firstly, a dual multimodal attention mechanism is designed to extract the explicit semantic information about the corrupted regions, which is done by comparing the descriptive text and complementary image areas through reciprocal attention. Secondly, an image-text matching loss is applied to maximize the semantic similarity of the generated image and the text. Experiments are conducted on two open datasets. Results show that the proposed TDANet model reaches new state-of-the-art on both quantitative and qualitative measures. Result analysis suggests that the generated images are consistent with the guidance text, enabling the generation of various results by providing different descriptions. Codes are available at https://github.com/idealwhite/TDANet.

*Corresponding author.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS CONCEPTS

• **Computing methodologies** → **Image processing**; • **Human-centered computing** → Natural language interfaces.

KEYWORDS

image inpainting, vision and language, image processing

ACM Reference Format:
Lisai Zhang, Qingcai Chen, Baotian Hu, Shuoran Jiang. 2020. Text-Guided Neural Image Inpainting. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Image inpainting is the task to generate visually realistic content in missing regions while keeping coherence [20]. It plays an important role in many digital image processing tasks, such as restoration of damaged paintings, photo editing, and image rendering [4]. There have been many methods proposed for generating semantically coherent content, such as integrating context features [11, 19], enhance convolution layers [18, 41]. and pluralistic completion [44] through probability framework.

A common assumption for inpainting models is that the missing area should have patterns similar to the remaining region. For example, diffusion-based [27] and patch-based [31] models use the remaining image to recover the missed regions. These models produce high-quality images, but usually usually fail in complicated scenes such as unique masks on objects or large holes [39]. In recent years, deep learning-based image inpainting methods have been presented to overcome this limitation [39]. An encoder-decoder inpainting structure with Generative Adversarial Networks (GAN) was first proposed by Pathak et al. [25]. Satoshi et al. [11] further improved the ability to encode the corrupted image and achieved photo-realistic results by adopting dilated convolutions and proposing global and local discriminators. However, the model cannot achieve high image appearance quality when filling irregular holes. Later works on neural image inpainting have achieved credible

generation quality on irregular masks from the perspective of improving convolution [20, 41] and integrating context features [19]. Although the appearance quality is constantly improving, it is still a challenge to generate accurate content for unique areas with patterns different than the remaining region.

Unique areas are hard to fill because there could be many semantically different solutions consistent with surrounding pixels. One of the directions for solving this problem is generating a diverse set of inpainting results. For example, a recent model PICNet [44] proposed a dual pipeline training architecture to learn distribution for the masked area. The model generates plausible pluralistic predictions for single masked input. However, the solution space for such methods is very large. Even if they generate all possible solutions, the desired result needs to be chosen from numerous outputs, which is not only time-consuming but also resource-intensive. Giving external guidance to inpainting models is another common approach to control the output and reduce the solution space. Some user-guided image inpainting approaches allow external guidance, such as a line [41] and edges [24], and exemplar [13]. However, these models provide only simple graphics tips and lack semantic diversity. What is more, the quality of user-guided inpainting largely depends on the quality of guidance, but suitable exemplar or drawing is sometimes difficult to obtain.

Since image content can be described as descriptive texts most of the time, it would be feasible for inpainting models to borrow information about the hole from text descriptions. For example, in Fig. 1, the color of the bird's belly is hard to be inferred only according to the remaining parts, but given a text description, the inpainting model will have a clear target. Some research on using text guidance already exists for other tasks, such as image generation [28, 32, 35, 42], These text to image generation works mainly focus on generating a complete image but do not take into account the constraints of image context. Image manipulation researchers also explored changing image content according to the text guidance [15, 23]. These models can change the pixels within the whole image area. Compared to text to image synthesis and text-guided manipulation, the text-guided inpainting task has more strict requirements. It requires identifying semantics that is complementary to existing image content, and generating coherent pixels at a fixed position.

In the text-guided inpainting task, the relationship between description and image samples is one-to-many: a semantic describes various samples that have the same meaning. It is natural to model such a relationship as probability distribution through the CVAE-like framework [6]. However, CVAE has been proven to have a disadvantage of grossly underestimating variances in the image completion scenario [44], which can be solved through the dual probabilistic structure. To use this structure in the text-guided inpainting task, we need to extend it to support the multimodal condition.

In this paper, we move one step further along user-guided inpainting and propose a novel model that fills the holes with the guidance of descriptive text. We design a novel dual multimodal attention mechanism to exploit the text features about the masked region by comparing text with the corrupted image and its counterpart. Then an image-text matching loss is adapted to regularize the similarity between text and model output. Experiments are conducted on two image caption datasets. The object boxes are used as the mask to unique areas. The inpainting results are evaluated in both qualitative and quantitative ways.

The main contributions of this paper are listed as follows:

- A text-guided dual attention model is proposed to complete image with the guidance of descriptive text. The combination of text and image provides richer semantics than only using the corrupted image.
- A novel inpainting scheme is presented that enables entering different texts to get pluralistic outputs.
- New state-of-the-art performances are reached by the proposed model on two public benchmarks.

2 RELATED WORK

Image inpainting Traditional diffusion-based or patch-based approaches [27, 31] fill missing regions by propagating neighbor information[2, 37] or copying from similar patches [8, 36] of the background based on low-level features. These methods work well for surface textures synthesis but often fail on non-stationary textural images [44]. To address the problem, Simakov et al. propose a bidirectional similarity synthesis approach [30] to better capture non-stationary information, but lead to high algorithm complexity. Recently, deep learning models are introduced to image inpainting that directly generates pixel values of the hole. Context encoders [25] use the encoder-decoder structure and conditional generative adversarial network to image inpainting tasks. Next, contextual attention [40] is proposed for capturing long-range spatial dependencies. To improve the performance on irregular masks, [18] proposes partial convolution where the mask is updated, and convolution weights are re-normalized with layers. However, it has limitations to capture detailed information of the masked area in deep layers. For this, [41] proposes to use gated convolution where a gate network learns the shape of mask in convolution. These approaches can produce only one result for each incomplete image. Thus, Zheng et al. [44] introduces a CVAE [3] based pluralistic image completion approach. During inference, the model obtains the distribution for the input, then samples various representation vectors from the distribution, and feeds these vectors to the decoder to get a variety of outputs. Our model builds upon this dual training structure and explores how it can be used for user-guided image inpainting.

User-guided Image inpainting Many user-guidance mechanisms are explored to enhance image inpainting systems, including dots or lines [1], structures [10], transformation or distortion [26], and exemplars [7, 13]. Some methods are also based on patch match algorithms, but use external image sources [5, 43] such as search engine [34]. Recent advances in conditional generative networks benefit user-guided inpainting and synthesis. Wang et al. [38] proposes to synthesize high-resolution photo-realistic images from semantic label maps using conditional generative adversarial networks. [41] extends this model to support user-guided inpainting with sketches. Compared with these schemes, image inpainting guided with text is challenging in two aspects: Firstly, image and text are heterogeneous; it is hard to transform image and text features to a shared space. What's more, the descriptive text usually contains redundant information, and the model must distinguish

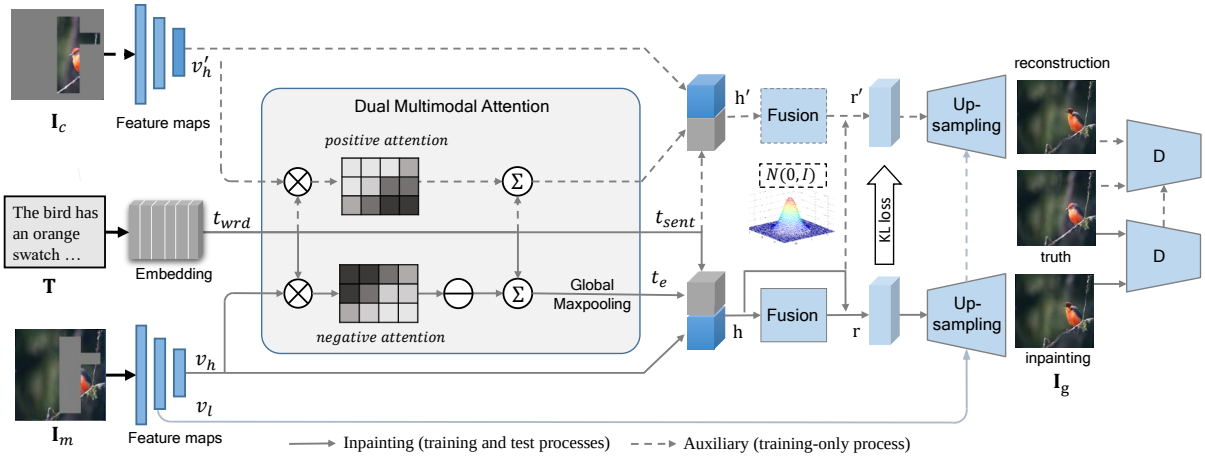


Figure 2: Architecture of the TDANet. The auxiliary path in dotted line is calculated only during training. The bold line is the inpainting path that infers the missing region from $\mathbf{I}_m$ and T. The two paths share weights except in fusion modules.

between the information about the corrupted region and the remaining parts.

Text-guided Image Synthesis and Manipulation The development of text to image synthesis brings the possibility of generating images based on text prior. The task is different from text-guided image inpainting and directly generates an image from the text description. Here we selectively review several related works. [28] first shows that the conditional GAN is capable of synthesizing plausible images from text descriptions. [42] stacks several GANs for text to image synthesis. However, their methods are conditioned on the global sentence vector, missing fine-grained word-level information. AttnGAN [35] develops word and sentence level matching networks to generate fine-grained images from text. More recent works [14, 16] exploit object and word-level information during synthesis through an attention mechanism. Compared with image synthesis task, the requirement for inpainting is more stringent: the generated content must coherent with remaining parts. Some image manipulation work [15, 23] explored text guidance; the method automatically edits an image and changes its content according to the text description. In manipulation task, the model decides which pixels to change, but inpainting requires to generate content for fixed position. Our research adopts the matching loss in AttnGAN to a new task, where the prior of generation is the combination of text and image, and generation target is a sub-region.

3 APPROACH

The proposed TDANet can be formulated as follows: given the masked input image $\mathbf{I}_m$ and descriptive text **T**, the model outputs the target image $\mathbf{I}_g$. The model uses the dual probabilistic structure and extend it to the multimodal condition. The overall structure of our model is shown in Figure 2. It composes of three components: Encoders for Image and Text, Dual multimodal Attention, and Inpainting Generation.

In the auxiliary path, the input $\mathbf{I}_c$ is the masked region of the original image. We use x' to denote the variable x is now calculated in the auxiliary path.

3.1 Preliminaries

The DAMSM loss [35] is composed of pre-trained networks that match the similarity of image and texts. By maximizing similarity score, these pre-trained networks calculate gradients as discriminators that enforce the generated image to follow the text description. In detail, the word-image similarity is defined as Eq. 1

$$S(I,T) = \log(\sum_{i=1}^{L-1} \exp(\gamma cos(I_i, T_i)))^{\frac{1}{\gamma}} \tag{1}$$

where I are the features of a generated image, T is the word embeddings, and L is the length of the sentence. For every batch of sentence-image pairs, the similarity score is computed as Eq. 2.

$$P(I,T) = \frac{\exp(\gamma_2 S(I_i, T_i))}{\sum_{j=1}^{B} \exp(\gamma_2 S(I_i, T_j))} \tag{2}$$

The word-image similarity could be optimized by minimizing the log probability of the score as Eq. 3.

$$DAMSM_w = -\sum_{i=1}^{B} (\log P(I,T) + \log P(T,I)) \tag{3}$$

The sentence-image similarity $DAMSM_s$ is calculated in the same way, except for redefining Eq. 1 as $S(I,T) = \cos(\bar{I}, \bar{T})$.

The short+long term attention [44] is proposed to effectively reconstruct the unmasked areas in the inpainting model. The network computes a weight map between encoder (long term) and decoder (short term) features through a self-attention network, and then weights and sums these feature maps. The generator concatenates the weighted feature maps and reconstructs the unmasked areas.

3.2 Encoders for Image and Text

In the image inpainting task, only the lost areas need to be inferred, and the remaining pixels could be reconstructed. We feed the input image to a 7-layer ResNet [9], and extract features from different layers. The feature map of the top layer is taken as the high-level representation v_h; the output of the second last layer is used as low-level features v_l . We use the high-level features to build the

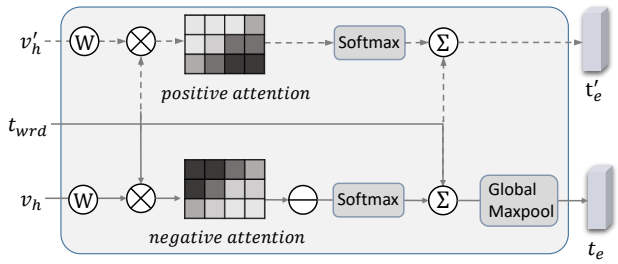


Figure 3: Detailed diagram of the Dual Multimodal Attention. The dotted path is auxiliary, while the bold one is the inpainting path. W denotes an 1x1 convolution layer.

prior condition of inference and feed the low-level feature v_l to the generator to reconstruct the unmasked areas. The image encoder in the auxiliary path shares the same weight with the one in the inpainting path.

As for the input sentence, a GRU network pretrained as [35] is used to compute a sequence of word representations t_{wrd} and a sentence representation t_{sent}. The hidden size of GRU network is 256. A sentence usually describes many parts of the image, so the key information about the missing region only exists in a subset of words. In order to use the word embedding to guide inpainting, the model must learn to extract this part of the information.

3.3 Dual Multimodal Attention Mechanism

The guidance provided by the text lies in the semantics about the missing region. Therefore, the model needs to compare the input text with the corrupted image, and highlight the words that do not match the image. In this matching task, there are only negative pairs that consists of $\mathbf{I}_m$ and $\mathbf{T}$. It would be hard to learn the match function if there are no positive pairs. Therefore, we propose a dual multimodal attention mechanism that takes advantage of the dual structure and uses $\mathbf{I}_c$ and $\mathbf{T}$ to constitute positive pairs. The mechanism extracts complementary features from the word embedding in two paths by reciprocal attention. Our basic assumption for this design is that $\mathbf{I}_m$ and $\mathbf{I}_c$ are complementary. The assumption is robust in the text-guided image inpainting scenario: if $\mathbf{I}_m$ and $\mathbf{I}_c$ are not complementary, we can solely use visual features v_h to infer the missing region, and do not need to specifically use the text features.

The detailed computation diagram of this mechanism is shown in Fig. 3. The visual representations of the two paths are firstly transformed through an 1x1 convolution. Then in the auxiliary path, the attention weights between v_h and word representations t_{wrd} are computed by multiplication. We keep the mask of the input image and apply it to the feature map on the same position. The computing can be formulated as Eq. 4:

$$s'_{i,j} = M'_i Q(v'_{hi})^T t_{wrdj} \tag{4}$$

where s denotes the attention map, M' is the binary mask (the value for masked pixels is 0 and elsewhere is 1), $Q(v'_h)$=Wv'_h, and W is the 1x1 convolution filter.

In the inpainting path, the attention weights between v_h and t_{wrd} is calculated to highlight the semantics about $\mathbf{I}_c$ as Eq. 5.

$$s_{i,j} = -Q(v_{hi})^T t_{wrdj} + M_i \tag{5}$$

The attention is calculated reciprocally to Eq. 4 by negating the multiplication results. Specifically, in M, the value for masked pixels is set to $-\infty$ and elsewhere to 0.

Attention maps of both paths are fed to softmax as Eq. 6 to get weights for the text embeddings.

$$\beta_{j,i} = \frac{\exp(s_{i,j})}{\sum_{i=1}^{N} \exp(s_{i,j})} \tag{6}$$

where N is the area of the feature map, β denotes the attention weights. According to the weights, word representations in both paths are weighted and summed up as Eq.7:

$$t_{ei} = \sum_{j=1}^{L} \beta_{i,j} t_{wrdj} \tag{7}$$

where L is the length of the sentence, t_e are the weighted word representations.

There is still a problem for t_e in the inpainting path now: after multiplying the attention weights, the features in t_e are distributed following position in v_h, so the values at the masked region in t_e are still zero, making it hard to be processed by the following convolutional layers. To handle this problem, we apply a global max-pooling on t_e. Since the specific location of missing semantic is unknown, we then uniformly replicate the max-pooling output.

3.4 Inpainting Generation

The inpainting generation part takes the image and text features as prior and predicts the original image. The features extracted from text and image are combined as multimodal hidden features h. Then, h is fed to a fusion network as the prior condition and projected to a latent space. We assume the latent space is a Gaussian distribution and use the fusion network to predict a group of parameters for the latent space. The fusion process in the inpainting path is formulated as Eq. 8

$$\mu, \sigma = F(h) \quad \text{where } h = [v_h; t_e; t_{sent}] \tag{8}$$

where μ and σ are mean and variance of the predicted Gaussian distribution, F denotes the fusion network, which consists of a 5-layer residual blocks with spectral normalization [22]. In the auxiliary path, distribution parameters μ' and σ' are calculated with the same process, but the fusion network F' does not share weights with F, because h and h' consists of different features.

Next, a multimodal representation r is obtained by sampling latent variables from the distribution and combine the variables with h through a residual connection as in Eq. 9,

$$r = h + Gaussian(\mu, \sigma) \tag{9}$$

where r is the multimodal representation. In the auxiliary path, the multimodal representation r' is also obtained based on the hidden representation h in the inpainting path as Eq. 10, because we need to keep the multimodal representation of these two paths homogeneous.

$$r' = h + Gaussian(\mu', \sigma') \tag{10}$$

Finally, the up-sampling networks produce a synthesized image $\mathbf{I}_g$ based on the multimodal representations on the two paths. Since image inpainting task does not require predicting the remaining areas, we feed the low-level image feature v_l to the generator through a high way path with short+long term attention mentioned in the preliminary section 3.1 to reconstruct the pixels of $\mathbf{I}_c$. The up-sampling network consists of 5-layer residual generator network shared by the two paths.

3.5 Optimization

3.5.1 Objective. In the TDANet, the inpainting path predicts the masked region based on $\mathbf{I}_m$ and $\mathbf{T}$, while the auxiliary path reconstructs the image because the information of the full images are available. Based on the visual features v_h, a complete distribution for $\mathbf{I}_m$ could be learnt in the auxiliary path, and be used to guide the distribution learning in the inpainting path.

To keep the homogeneity between the latent space in the auxiliary path and inpainting path, we also feed text features to the auxiliary path, and reconstruct the image based on multimodal prior h'. The conditional variational lower bound of the auxiliary path is formulated as Eq. 11:

$$\begin{aligned}\log p(\mathbf{I}_c|h') \geq & -\mathrm{KL}(q_\psi(\mathbf{z}|\mathbf{I}_c, h')||p_\phi(\mathbf{z}|h')) \\ & + \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{I}_c,h')}[\log p_\theta(\mathbf{I}_c|\mathbf{z})]\end{aligned} \tag{11}$$

where $\mathbf{z}$ is the latent vector, $q_\psi(\cdot|\cdot)$ denotes posterior sampling function, $p_\phi(\cdot|\cdot)$ denotes the conditional prior, $p_\theta(\cdot|\cdot)$ denotes the likelihood, with ψ, ϕ, and θ being the deep network parameters of their corresponding functions.

In the inpainting path, the prior consists of the features from $\mathbf{I}_c$ and $\mathbf{T}$. The lower bound is formulated as Eq. 12:

$$\begin{aligned}\log p(\mathbf{I}_c|h) \geq & -\mathrm{KL}(q_\psi(\mathbf{z}|\mathbf{I}_c, h)||p_\phi(\mathbf{z}|h)) \\ & + \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{I}_c,h)}[\log p_\theta(\mathbf{I}_c|\mathbf{z})]\end{aligned} \tag{12}$$

The inpainting quality could be optimized by maximizing the variation lower bound. However, $q_\psi(\mathbf{z}|\mathbf{I}_c, h)$ can not be calculated directly because $\mathbf{I}_c$ is not available in the inpainting path. To solve this problem, we assume that r could approximate to r' after optimizing inpainting path encoders, so we have $q_\psi(\mathbf{z}|\mathbf{I}_c, h') \approx q_\psi(\mathbf{z}|\mathbf{I}_c, h)$, and Eq. 12 is updated as Eq.13:

$$\begin{aligned}\log p(\mathbf{I}_c|h) \geq & -\mathrm{KL}(q_\psi(\mathbf{z}|\mathbf{I}_c, h')||p_\phi(\mathbf{z}|h)) \\ & + \mathbb{E}_{q_\psi(\mathbf{z}|\mathbf{I}_c,h)}[\log p_\theta(\mathbf{I}_c|\mathbf{z})]\end{aligned} \tag{13}$$

The parameters of the TDANet are learnt by maximizing these two lower bound and minimize the distance between $q_\psi(\mathbf{z}|\mathbf{I}_c, h')$ and $q_\psi(\mathbf{z}|\mathbf{I}_c, h)$.

3.5.2 Loss Function. We assumed in section 3.4 that the latent vector follows a Gaussian distribution. To optimize the variation lower bound and enforce the smoothness of the conditioning manifold space [6], we maximize the Kullback-Leibler divergence term between and the standard Gaussian distribution. For the auxiliary path, the distribution loss is formulated as Eq.14:

$$\mathcal{L}'_{KL} = -KL(q_\psi(\mathbf{z}|\mathbf{I}_c, h')||\mathcal{N}(0, 1)) \tag{14}$$

As for the inpainting path, we steer the conditional prior to close to the auxiliary path posterior, which can be formulated as Eq. 15:

$$\mathcal{L}_{KL} = -KL(q_\psi(\mathbf{z}|\mathbf{I}_c, h')||p_\phi(\mathbf{z}|h)) \tag{15}$$

Dataset	CUB	COCO
Mask Area (Avg)	34.59%	19.62%
Vocabulary size	5,450	27,297
Caption length (Avg)	15.23	10.45
Objects per image (Avg)	1	7.3
Captions per image	10	5

Table 1: Statistics for object mask and related properties on CUB and COCO dataset. The average mask area of COCO is calculated with the max box.

Next, the likelihood term $p_\theta(\mathbf{I}_c|\mathbf{z})$ need to be optimized. The likelihood term could be interpreted from both appearance and semantic aspects.

If $\mathbf{I}_g$ is close to $\mathbf{I}$ in appearance distance, the likelihood of generating $\mathbf{I}_c$ is improved. For image appearance quality, we incorporate reconstruction and adversarial losses. Three loss functions are applied as Eq. 16.

$$\mathcal{L}_I = ||\mathbf{I} - \mathbf{I}_g||_1 + ||D_1(\mathbf{I}) - D_1(\mathbf{I}_g)||_2 + [D_2(\mathbf{I}_g) - 1]^2 \tag{16}$$

The first term matches the per-pixel distance through L_1 distance. The second term is mean feature match loss [6], where D_1 is a discriminator that returns a representation vector from the final layer. In the third term, D_2 indicates whether $\mathbf{I}_g$ is a real word image, the term is based on the loss in LSGAN [21], which has been proven [44] to perform better than the original GAN loss in the inpainting scenario. The networks of D_1 and D_2 are 5-layer ResNet.

To refine the semantic relation of text and image output, we adapt the DAMSM loss to our model. The match networks extract features from the generated image $\mathbf{I}_g$ through a text-image attention network and compare these features with text features. We follow the setup in [35] and set γ to 5 and formulate the loss term in Eq. 17.

$$\mathcal{L}_T = DAMSM(t_{wrd}, \mathbf{I}_g) \tag{17}$$

The loss is calculated in both paths with shared matching networks.

Finally, the total loss function could be formulated as Eq. 18:

$$\mathcal{L} = \lambda_{KL}(\mathcal{L}'_{KL} + \mathcal{L}_{KL}) + \lambda_I(\mathcal{L}_I + \mathcal{L}'_I) + \lambda_T(\mathcal{L}_T + \mathcal{L}'_T) \tag{18}$$

where $\mathcal{L}_{KL}$ regularizes KL divergence of prior and posterior distribution, $\mathcal{L}_I$ and $\mathcal{L}_T$ maximize the expectation term from image quality and text-image semantic similarity perspectives.

4 EXPERIMENTS

4.1 Datasets

4.1.1 Images and Descriptions. The proposed method is evaluated on image captioning datasets CUB [33] and COCO [17], with their original train, test, and validation split. The captions are used as a description for inpainting. Compared with CUB, COCO captions have a larger vocabulary but shorter and fewer sentences for each image. The properties of the two datasets are given in Table 1. COCO contains more instances and scenes than CUB, so learning to complete the images of coco is more challenging than CUB.

4.1.2 Center Mask. Following common inpainting evaluation setting [19, 20, 41, 44], a square mask taking 50% area is constructed at the center of every image. The center mask is wildly used to compare inpainting models, but could not test their robustness at unique areas.

Dataset	Model	ℓ_1^-(%)			PSNR$^+$			TV loss$^-$(%)			SSIM$^+$(%)		
		center	object	||Δ||	center	object	||Δ||	center	object	||Δ||	center	object	||Δ||
CUB	CSA [19]	3.99	6.42	2.43	20.79	19.13	1.66	3.64	4.33	0.69	82.69	72.65	10.4
	PICNet [44]	3.65	7.28	3.63	20.96	18.80	2.16	3.71	4.12	0.41	84.51	70.78	13.73
	TDANet	**3.53**	**4.80**	**1.27**	**21.30**	**20.89**	**0.41**	**3.55**	**3.34**	**0.21**	**84.63**	**79.16**	**5.47**
COCO	CSA [19]	5.07	8.78	3.71	20.07	19.23	0.84	4.19	4.86	0.67	83.21	75.22	7.99
	PICNet [44]	5.53	9.21	3.68	19.57	18.73	1.02	4.51	4.97	0.46	81.78	74.44	7.34
	TDANet	**4.08**	**7.48**	**3.40**	**21.31**	**20.57**	**0.74**	**4.54**	**4.20**	**0.34**	**83.87**	**76.78**	**7.09**

Table 2: Quantitative comparison with the state-of-the-art on CUB and COCO dataset. ||Δ|| is the performance difference between the object mask and the center mask, showing the robustness of the method at different mask positions. $^-$ lower is better, $^+$ higher is better.

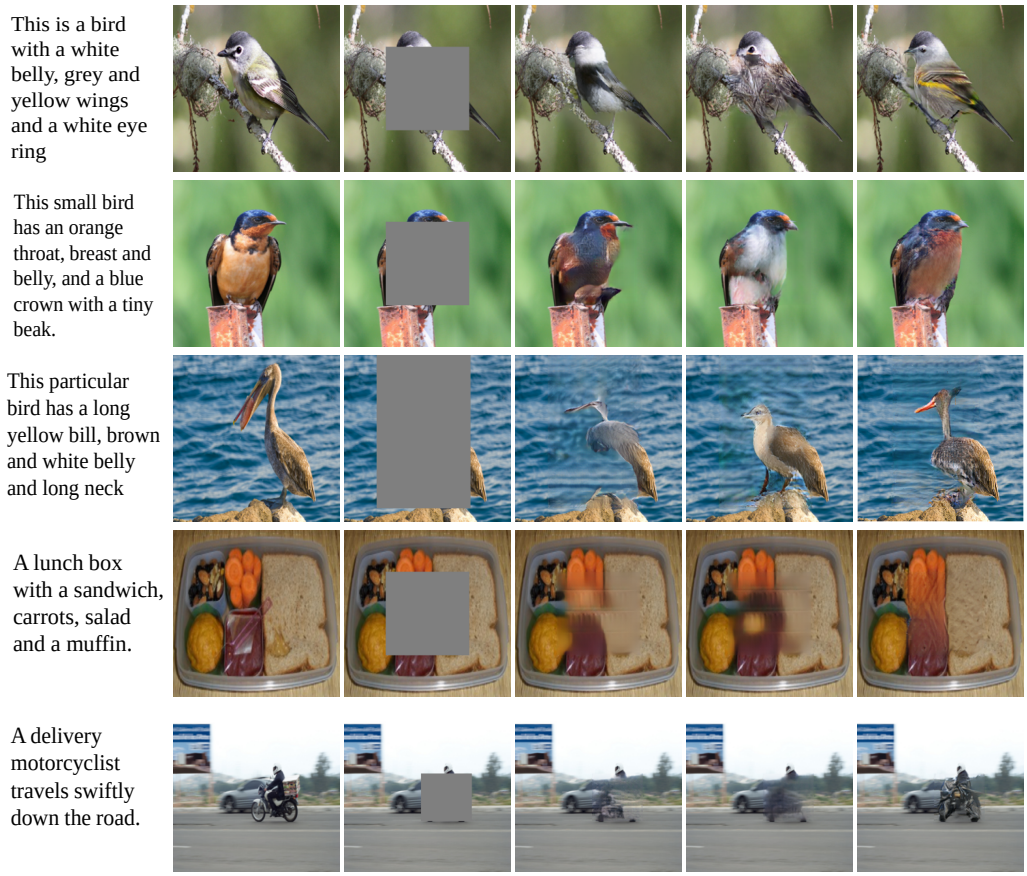


Figure 4: Qualitative comparisons on CUB and COCO validation sets. (Best viewed with zoom-in)

4.1.3 Object Mask. Object regions are important to an image but are hard to restore. These regions are ideal test environment for unique area inpainting evaluation but have not been used yet. To evaluate our model, we construct an object mask from object detection boxes. The COCO dataset provides dense object boxes for each image, we only select the one with the largest area, because using all masks will sometimes remove most pixels of an image and break the content. The object masks are various in size depending on the size of the objects.

4.2 Implementation Details

For images in both CUB and COCO datasets, the training images are resized to make their minimal length/width as 256 and crop the sub-image of size 256x256 at the center. Networks are initialized with orthogonal initialization [29] and trained end-to-end with a learning rate of 10^{-4} on Adam optimizer [12]. The image-text matching networks are pre-trained as in [35] on CUB and COCO, respectively. During training, the weights of loss function terms are set as $\lambda_{KL} = \lambda_I = 20$, $\lambda_T = 0.1$. The maximum length of text sequence is set to 128 with zero padding. Experiments are conducted

Text 1: "This bird has a black crown and tiny beak. the breast, throat, and nape are yellow. the wings are black with white wingbars."
Text 2: "The bird has brown throat, white breast, belly and abdomen, gray wings with two wingbars."

Figure 5: Qualitative cases of the ablation study.

on Ubuntu 18.04 system, with i7-9700K 3.70GHz CPU and 11G NVIDIA RTX2080Ti GPU. We compare the proposed TDANet with two state-of-the-art image inpainting methods PICNet [44] and CSA [19] based on their official source codes [1].

4.3 Quantitative Results

Inpainting quality depends on the authenticity of the filled area and its continuity with the surrounding images. Existing quantitative metrics for inpainting models could only evaluate these characteristics roughly [40, 41]. To compare with other models, we select commonly used mean ℓ_1 loss, peak signal-to-noise ratio (PSNR), total variation (TV), and Structural Similarity (SSIM) for quantitative comparison. We will further discuss the semantic consistency in other experiments.

From the results in Table 2, the proposed TDANet outperforms compared models in all measures. The performance improvement on the object mask is more obvious than the center mask. Comparing the improvements on the two datasets, TDANet improves more on CUB than COCO, which qualifies our expectation that combining image and caption on COCO is more challenging. It is worth noting that the performance of all evaluated models decreases on object masked samples, while our model has the least performance degradation, proving the robustness in such unique areas.

4.4 Qualitative Results

Both inpainting quality and semantic consistency are evaluated in qualitative comparison. Fig. 4 shows the qualitative results on the CUB and COCO validation set. As shown in the figure, the results of all the three models on the CUB dataset are consistent with the surrounding areas but compared with the ground truth, the CSA and PICNet outputs fail to recover some special characteristics, such as the belly color and neck shape. Through careful observation, it can be found that the content filled by CSA and PICNet has similar characteristics to the neighbor pixels such as color and texture. In other words, these models are filling unique areas by borrow features from their surroundings. In comparison, the images generated by the TDANet produce these unique features as mentioned in the text, so the results look semantically consistent with the original image. On the challenging COCO dataset, all these three models are imperfect, but our model obtains a better appearance quality compared with other models.

[1] https://github.com/KumapowerLIU/CSA-inpainting
https://github.com/lyndonzheng/Pluralistic-Inpainting

Model	Naturalness	Semantic Consistency
Ground Truth	1.208	1.363
CSA	2.981	3.060
PICNet	3.178	3.469
TDANet	**2.531**	**2.106**

Table 3: Numerical ranking score of the user study. The lower the score, the better the performance.

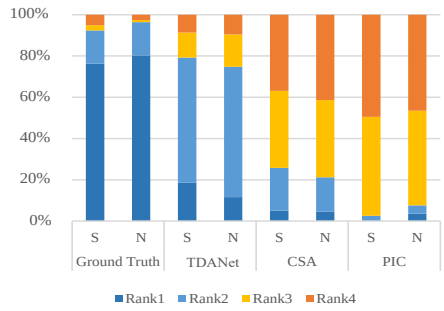


Figure 6: Ranking score distribution of the user study. "S" means semantic consistency score, "N" means naturalness.

4.5 User Study

A ranking game was designed to further quantify the qualitative comparison from the human perspective. We randomly collected 100 images in center masks from the CUB validation dataset (see details in supplementary files). For each image, a tuple consisting of (1) Ground truth, (2) Our model, (3) PICNet, (4) CSA model is prepared. We randomly shuffled the four samples in each tuple and recruited 20 volunteers to rank them according to naturalness and the semantic consistency with the text description. After the test, we computed the average of the ranking score for the four groups.

Model	ℓ_1^-	PSNR$^+$	TV loss$^-$	SSIM$^+$
TDANet	4.80	20.89	3.34	79.16
w/o match loss	4.81	20.93	3.48	78.58
w/o maxpool	4.71	20.83	3.48	78.85
w/o multimodal attention	5.77	20.14	3.78	74.31

Table 4: Numerical results of the ablation study.

The results are shown in Table 3. According to the results, our model performs better than other models in terms of realistic, and significantly higher in semantic consistency. We also counted the rank score distribution of each model in Fig. 6. Compared with other models, TDANet results ranks first and second more often than the other two models, and third and fourth less often. We also found that sometimes TDANet could mislead the tester and get a better rank than the original image. We also noticed that TDANet gets ranks better in semantic consistency than naturalness, which means sometimes the model grasps the key information from the text but can not generate the content perfectly due to the limitation of the generator.

4.6 Ablation Study

Four models without each component were trained using the same super parameters and epochs. As shown in Table 4, w/o the auxiliary path and dual multimodal attention lead to the most obvious performance decrease on every metric. w/o the global max-pooling layer and replication of text features mainly lead to appearance metrics decrease, which is reflected by TV loss and SSIM. We have noticed that removing the match loss leads to appearance metrics decrease while pixel metrics increases.

To further inspect the effect of the components, we compared their result qualitatively; the cases are given in Fig 5. The quantitative ablation comparison shows that although w/o the match loss leads to higher pixel metrics, its output looks worse and fails to allocate characters precisely.

4.7 Controllability Evaluation and Manipulation Extension

The text guidance not only benefits inpainting quality but also allows changing the model output by providing various descriptions. The interactive manner enables more application scenarios, such as manipulation. Here we show the controllability and extension to image manipulation at the same time. The manipulation has three steps: First, select the region to be changed and mask it. Second, write a sentence describing the desired image after manipulation. Third, feed the masked image and sentence to the TDANet and get the result. For each image, we present two results when two different sentences are entered.

Figure 7 shows the image manipulation results. The first group is color variation by masking the belly and giving sentences to describe the bird with different colors. The second group shows the edit of different parts by providing sentences describing different organs. In the third case, we selected an area between blue and red colors and command the model to fill in one of the neighboring colors. The result shows the output was not affected by the color of adjacent areas. The experiment presents a new possible formulation

Input Mask Result (a) Result (b)

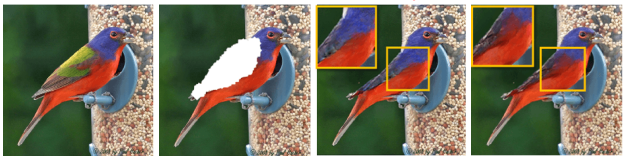
Text (a): "The bird has ***black and white spotted breast*** *a red hat and grey wings."*

Text (b): "The bird has a purple breast a red hat and grey wings."

Input Mask Result (a) Result (b)

Text (a): "A bird with a white breast and a brown crown."

Text (b): "a bird with a white breast blue head and red eyes. "



Input Mask Result (a) Result (b)

Text (a): "A distinct red bird with a blue head blue wings and a red eyering."

Text (b): "A distinct red bird with a blue head and red wings."

Figure 7: TDANet with different guidance texts.

of image manipulation, and shows that pluralistic results could be generated by controlling the guidance text.

5 CONCLUSION

We have presented a text-guided inpainting scheme that combines the features of the image and descriptive text as the generation condition. We proposed a dual multimodal attention mechanism to exploit semantic features about the masked region from the descriptive text. What's more, we introduced an image-text matching based loss to improve the semantic consistency between inpainting output and the text. The experimental results demonstrated that the proposed TDANet outperformed compared models in subjective and objective comparisons, and the model outputs are semantically consistent with the guidance text.

Several future directions and improvements could be considered. Our main objective was to show the potential of text-guided image inpainting. We found that with simple encoders, the semantics were well extracted; future work might consider more complex architectures to refine the representations. Performance on the COCO dataset is still limited. External visual knowledge about concepts and data augmentation will improve inpainting quality.

ACKNOWLEDGMENTS

We would like to thank Joanna Siebert for her helpful feedback. This work is supported by Natural Science Foundation of China (Grant No. 61872113), Strategic Emerging Industry Development Special Funds of Shenzhen (Grant No.XMHT20190108009), the Tencent Group and Science and Technology Planning Project of Shenzhen (Grant No.JCYJ20190806112210067).

REFERENCES

[1] Michael Ashikhmin. 2001. Synthesizing natural textures. In *Proceedings of the 2001 Symposium on Interactive 3D Graphics, SI3D 2001, Chapel Hill, NC, USA, March 26-29, 2001*. 217–226.

[2] Coloma Ballester, Marcelo Bertalmío, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. 2001. Filling-in by joint interpolation of vector fields and gray levels. *IEEE Trans. Image Process.* 10, 8 (2001), 1200–1211.

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. 2017. CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. 2764–2773.

[4] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 417–424.

[5] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. 2004. Region filling and object removal by exemplar-based image inpainting. *IEEE Trans. Image Process.* 13, 9 (2004), 1200–1212.

[6] Carl Doersch. 2016. Tutorial on Variational Autoencoders. *CoRR* abs/1606.05908 (2016). arXiv:1606.05908

[7] James Hays and Alexei A. Efros. 2007. Scene completion using millions of photographs. *ACM Trans. Graph.* 26, 3 (2007), 4.

[8] Kaiming He and Jian Sun. 2014. Image Completion Approaches Using the Statistics of Similar Patches. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 12 (2014), 2423–2435.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 770–778.

[10] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. 2014. Image completion using planar structure guidance. *ACM Trans. Graph.* 33, 4 (2014), 129:1–129:10.

[11] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Trans. Graph.* 36, 4 (2017), 107:1–107:14.

[12] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[13] Vivek Kwatra, Irfan A. Essa, Aaron F. Bobick, and Nipun Kwatra. 2005. Texture optimization for example-based synthesis. *ACM Trans. Graph.* 24, 3 (2005), 795–802.

[14] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. 2019. Controllable Text-to-Image Generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*. 2063–2073.

[15] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr. 2019. ManiGAN: Text-Guided Image Manipulation. *CoRR* abs/1912.06203 (2019).

[16] Wenbo Li, Pengchuan Zhang, Lei Zhang, Qiuyuan Huang, Xiaodong He, Siwei Lyu, and Jianfeng Gao. 2019. Object-Driven Text-To-Image Synthesis via Adversarial Training. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 12174–12182.

[17] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, Vol. 8693. 740–755.

[18] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image Inpainting for Irregular Holes Using Partial Convolutions. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, Vol. 11215. 89–105.

[19] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. 2019. 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*. 4169–4178.

[20] 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 *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*. 3123–3129.

[21] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. 2019. On the Effectiveness of Least Squares Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 12 (2019), 2947–2960.

[22] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

[23] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. [n. d.]. Text-Adaptive Generative Adversarial Networks: Manipulating Images with Natural Language. In *Advances in Neural Information Processing Systems*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 42–51.

[24] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. 2019. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. *CoRR* abs/1901.00212 (2019).

[25] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. 2016. 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*. IEEE Computer Society, 2536–2544.

[26] Darko Pavic, Volker Schönefeld, and Leif Kobbelt. 2006. Interactive image completion with perspective correction. *The Visual Computer* 22, 9-11 (2006), 671–681.

[27] Chuan Qin, Shuozhong Wang, and Xinpeng Zhang. 2012. Simultaneous inpainting for image structure and texture using anisotropic heat transfer model. *Multim. Tools Appl.* 56, 3 (2012), 469–483.

[28] Scott E. Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, Vol. 48. 1060–1069.

[29] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

[30] Yuhang Song, Chao Yang, Zhe L. Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C.-C. Jay Kuo. 2018. Contextual-Based Image Inpainting: Infer, Match, and Translate. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, Vol. 11206. 3–18.

[31] Harry Strange, Ian Scott, and Reyer Zwiggelaar. 2014. Myofibre segmentation in H&E stained adult skeletal muscle images using coherence-enhancing diffusion filtering. *BMC Medical Imaging* 14 (2014), 38.

[32] Aäron van den Oord, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves. 2016. Conditional Image Generation with PixelCNN Decoders. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 4790–4798.

[33] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).

[34] Oliver Whyte, Josef Sivic, and Andrew Zisserman. 2009. Get Out of my Picture! Internet-based Inpainting. In *British Machine Vision Conference, BMVC 2009, London, UK, September 7-10, 2009. Proceedings*. 1–11.

[35] 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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 1316–1324.

[36] Zongben Xu and Jian Sun. 2010. Image Inpainting by Patch Propagation Using Patch Sparsity. *IEEE Trans. Image Process.* 19, 5 (2010), 1153–1165.

[37] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. 2018. Shift-Net: Image Inpainting via Deep Feature Rearrangement. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, Vol. 11218. 3–19.

[38] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. 2017. High-Resolution Image Inpainting Using Multi-scale Neural Patch Synthesis. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 4076–4084.

[39] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. 2017. Semantic Image Inpainting with Deep Generative Models. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 6882–6890.

[40] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. Generative Image Inpainting With Contextual Attention. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. 5505–5514.

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2019. Free-Form Image Inpainting With Gated Convolution. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. 4470–4479.

[42] Han Zhang, Tao Xu, and Hongsheng Li. 2017. 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*. 5908–5916.

[43] Yinan Zhao, Brian L. Price, Scott Cohen, and Danna Gurari. 2019. Guided Image Inpainting: Replacing an Image Region by Pulling Content From Another Image. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*. IEEE, 1514–1523.

[44] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2019. Pluralistic Image Completion. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. 1438–1447.