Neural Image Inpainting Guided with Descriptive Text

Lisai Zhang, Qingcai Chen, Baotian Hu and Shuoran Jiang
Harbin Institute of Technology, Shenzhen
lisaizhang@foxmail.com, qingcai.chen@hit.edu.cn, {baotianchina,shaunbysn}@gmail.com

Abstract
Neural image inpainting has achieved promising performance in generating semantically plausible content. Most of the recent works mainly focus on inpainting images depending on vision information, while neglecting the semantic information implied in human languages. To acquire more semantically accurate inpainting images, this paper proposes a novel inpainting model named Neural Image Inpainting Guided with Descriptive Text (NIGDT). First, a dual multi-modal attention mechanism is designed to extract the explicit semantic information about corrupted regions. The mechanism is trained to combine the descriptive text and two complementary images through reciprocal attention maps. Second, an image-text matching loss is designed to enforce the model output following the descriptive text. Its goal is to maximize the semantic similarity of the generated image and the text. Finally, experiments are conducted on two open datasets with captions. Experimental results show that the proposed NIGDT model outperforms all compared models on both quantitative and qualitative comparison. The results also demonstrate that the proposed model can generate images consistent with the guidance text, which provides a flexible way for user-guided inpainting. Our systems and code will be released soon.

1 Introduction
Image inpainting plays an important role in many tasks, such as restoration of damaged paintings, photo editing and image rendering [Bertalmio et al., 2000]. It requires generating visually realistic content in missing regions while keeping coherence [Ma et al., 2019]. Recently, there have been many methods proposed for generating semantically plausible and diverse contents, such as integrating context [Liu et al., 2019] and estimating the distribution of missed regions [Zheng et al., 2019].

Most image inpainting works are focused on the input image itself, which is based on the assumption that the missing area should have similar patterns with the background region. Based on this assumption, diffusion-based [Bertalmio et al., 2000] and patch-based [J. Weickert, 1999] models are proposed to use the remaining image to recover the missed regions. These models produce high quality images, but usually make critical failures in complicated scenes such as mask on objects or large holes [Yeh et al., 2017]. In recent years, deep learning based image inpainting methods have been presented to overcome this limitation [Yeh et al., 2017]. [Pathak et al., 2016] firstly proposed an encoder-decoder structure with Generative Adversarial Networks (GAN). To improve the ability of encoding the input image, [Iizuka et al., 2017] adopted dilated convolutions and proposed global and local discriminators and generate photo-realistic results. However, the model has limitations of filling irregular holes. The latest works on neural image inpainting have achieved credible generation quality on irregular masks from the perspective of improving convolution [Yu et al., 2019; Ma et al., 2019] and integrating context features [Liu et al., 2019]. Generating a diverse set of inpainting results is also an important research direction. The latest model is the PICNet [Zheng et al., 2019], which proposed a dual pipeline training architecture to learn distribution for the masked area. The model can generate diverse plausible content for a single masked input.

Giving external guidance to inpainting models is a commonly used solution for controlling generation. Some user-guided image inpainting approaches allow external guidance, such as line [Yu et al., 2019] and edges [Nazeri et al., 2019] and exemplar [Vivek Kwatra and Kwatra, 2005] but these guidelines are limited to specific graphics tips and lacks semantic diversity.

Since image semantics can be described as descriptive text in most cases, it would be plausible if inpainting models borrow necessary semantic information from text descriptions. For example, in Fig. 1, when a text description is given, the
target of the model becomes clear and the model output becomes controllable. However, to the best of our knowledge, there is still no work incorporating text into the image inpainting model to guide the generation.

In this paper, we focus on incorporating descriptive text into the neural image inpainting model. The main contributions of this paper are given below:

1) We propose a novel neural image inpainting model that fills the holes of an image with the guidance of descriptive text.

2) We propose a novel dual multi-modal attention mechanism to exploit the semantically features about the masked region from the descriptive text.

3) We design an image-text matching loss to regularize the similarity of text and model output and enforce the generated image following the descriptive text.

Experimental results of quantitative and qualitative comparison show that the proposed NIGDT model outperforms state-of-the-art models on open datasets with captions, and the generated results are semantically consistent with the guidance text.

2 Related Work

Our research builds on previous works in the field of image inpainting and text to image synthesis.

Traditional diffusion-based or patch-based approaches [Bertalmio et al., 2002; J. Weickert, 1999] fill in missing regions by propagating neighboring information or copying from similar patches of the background based on low-level features. These methods work well for surface textures synthesis but often fail on non-stationary textural images. To address the problem, Simakov et al. propose a bidirectional similarity synthesis approach [Yuhang Song and Jay, 2018] 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 [Pathak et al., 2016] firstly use the encoder-decoder structure and conditional generative adversarial network to image inpainting task. Further, contextual attention [Yu et al., 2018] is proposed for capturing long-range spatial dependencies. To improve the performance on irregular masks, partial convolution [Liu et al., 2018] is proposed 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, [Yu et al., 2019] presents to use gated convolution for free-form image inpainting 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., 2019] introduces a CVAE based pluralistic image completion approach. During inference, the model firstly obtains the distribution for the input, then sample various representation vectors from the distribution, and feed these vectors to the decoder to get a variety of outputs.

Many user-guided inpainting methods are explored to enhance image inpainting systems, including dots or lines, structures, transformation or distortion, and exemplars concluded in [Yu et al., 2019]. Recent advances in conditional generative networks empower user-guided inpainting and synthesis. Wang et al. [Ting-Chun Wang and Catanzaro, 2018] propose to synthesize high-resolution photo-realistic images from semantic label maps using conditional generative adversarial networks. [Yu et al., 2019] extend their model to support user-guided inpainting with sketches. Another related work [Li et al., 2019b] explores an image manipulation approach through text guidance. However, image inpainting with text is still challenging in two aspects: Firstly, image and text are heterogeneous, it is hard to transform image and text features to share space. What’s more, the descriptive text usually contains redundant information, the model must distinguish the information about the corrupted region form the text.

Text to image synthesis directly generates an image from the text description. Here we selectively review several related work. [S. Reed and Lee, 2016a] first showed that the conditional GAN was capable of synthesizing plausible images from text descriptions. [S. Reed and Lee, 2016b] stacked several GANs for text-to-image synthesis. However, their methods are conditioned on the global sentence vector, missing fine-grained word-level information. AttnGAN [Xu et al., 2018] develops word and sentence level matching network to generate fine-grained images from text. [Li et al., 2019a] proposed to exploit word-level information during synthesis through attention mechanism.

3 Approach

Our proposed NIGDT can be formulated as follows: given the masked input image $I_m$ and descriptive text $T$ the model output the target image $I_g$. The overall structure of our model is shown in Figure 2. It composes of three parts: Encoders for Image and Text, Dual multi-modal Attention and Inpainting Generation.

3.1 Encoders for Image and Text

The input image extract image representations with a 7-layer resnet. To obtain the image representations of low-level and high-level, we take the outputs of the last two layers, where $v_h$ is the high-level representation output by the last layer, $v_l$ is the output of the second last layer. $v_l$ will be fed directly to the generator to reconstruct the unmasked areas.

As for the input sentence, we use a text encoder to compute a sequence of word representations and a sentence representation, where $w_{word}$ are word representation vectors, $w_{sent}$ is sentence representation vectors. The text encoder is a GRU network with 256 hidden size.

3.2 Dual Multi-modal Attention Mechanism

The guidance provided by the text lies on the semantics about the missing region. However, it is hard to locate these semantics precisely without the image of missing region. Therefore, we propose a dual multi-modal attention mechanism to capture the relation between query and complementary contents, as shown in Fig. 2. The goal of dual multi-modal attention mechanism is to extract the semantic information about the masked area from the text. The key insight is: text is necessary only if $I_m$ and $I_s$ are complement, else we can solely use visual features to predict the whole image.
We construct a dual pair with $I_m$ and $I_c$, where $I_c$ is available only during training. It is called dual pair because we calculate the attention weights on these two inputs reciprocally. To formulate, we use $x^p$ to denote the variable $x$ only appears in the positive attention path, and $x^n$ for the negative attention path.

The high-level visual representation $v_h^p$ of $I_c$ is extracted through the same convolutional network as $v_h^n$, and then transform the visual representation with 1x1 convolution. As the dotted line path, we then compute the attention weights between $v_h^p$ and word representations $t_{wrd}$. When computing the attention weights we keep the mask from the input and apply it to the feature map on the same region. The computing is formulated as Eq. 1:

$$s_{i,j}^p = M_p Q(v_h^p)^T t_{wrdj}$$

where $Q(v_h) = W v_h$, and $W$ is an 1x1 convolution filter, $M_p$ is the binary mask (the value for masked pixels is 0 and elsewhere is 1).

As the bold line path in Fig 2, we then calculate attention weights between $v_h^n$ and $t_{wrd}$ reciprocally as Eq. 2.

$$s_{i,j}^n = -Q(v_h^n)^T t_{wrdj} + M^n$$

where $M^n$ is the mask on $v_h^n$. Specially, in $M^n$ the value for masked pixels is $-\infty$ and else where is 0.

Attention weights on both paths are fed to softmax as Eq 3

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

where $N$ is the area of the feature map, $\beta$ denotes the reciprocal attention map. The word representations are then weighted by multiplying the attention map as Eq.4:

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

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

There is still a problem for the negative attention path: vectors in $t_e$ are distributed following the content in $v_h$, so the values of the masked region in $t_e$ are zero, making it hard to be decoded. To handle this problem, we apply a global maxpooling to $t_e$. Because the specific location of missing semantic is unknown, we then uniformly replicate the max-pooling output. The calculation is formulated as Eq. 5

$$t_{ej} = \max_{1<i<N} \hat{t}_{ei}$$

where $N$ is the area of feature map, $t_e$ is the extracted word representation.

After extracting word representation through the multimodal attention, the image feature and extracted word representation are concatenated and fed to a fusion network to generate parameters for a Gaussian distribution, formulated as Eq. 7

$$h = [v_h, t_e]$$

$$\mu, \sigma = F(h)$$

where $\mu$ and $\sigma$ are mean and variance of the predicted Gaussian distribution, $F$ denotes the fusion network, which consists of 5-layer residual blocks. Finally, we combine the multi-modal hidden representations and sampled representation with a residual connection as in Eq. 8.

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

where $r$ is the multi-modal representation.

### 3.3 Inpainting Generation

The output image is generated based on multi-modal representation $r$ through a 5-layer residual network with spectral normalization [Miyato T, 2018]. We feed the low-level image feature $v_l$ to the generator through a high way path based on short long term attention as [Zheng et al., 2019]. The short long term attention firstly computes a weight map through self-attention on the decoder feature map, and then obtain short and long term attention features by multiply attention.
A lunch box with a sandwich, carrots, salad and a muffin.

This particular bird has a long yellow bill, brown and white belly and long neck.

This is a bird with a white belly, grey and yellow wings and a white eye ring.

This particular bird has a long yellow bill, brown and white belly and long neck.

This small bird has an orange throat, breast and belly, and a blue crown with a tiny beak.

This small bird has a white belly, black breast, neck, and crown with a tiny pointy bill.

The bird has brown throat, white breast, belly and abdomen, gray wings with two wingbars.

This is a bird with a white belly, grey and yellow wings and a white eye ring.

A blue and black bird which has metallic or florescent hues with a jet black head and dark wings.

A lunch box with a sandwich, carrots, salad and a muffin.

This is a bird with a white belly, black breast, neck, and crown with a tiny pointy bill.

This bird has a white belly, black breast, neck, and crown with a tiny pointy bill.
However, $q(\psi(z|I_c, h^1))$ cannot be calculated directly because $I_c$ is not available in generative path. We assume that $h^1$ is approximate to $h^0$ after optimizing encoders, and have $q(\psi(z|I_c, h^1)) \approx q(\psi(z|I_c, h^0))$, so Eq. 13 is updated as Eq.14:

$$
\log p(I_c|h^0) \geq -KL(q(\psi(z|I_c, h^1)||\mathcal{N}(0, 1)) + \mathbb{E}_{q(\psi(z|I_c, h^1))}[\log p_\theta(I_c|z)]
$$

(14)

The parameters of the NIGDT model are learnt by maximizing these two lower bound. Assuming prior distributions of $p_\theta(z|h^0)$ and $p_\theta(z|h^0)$ are Gaussian, for the reconstructive path, the distribution loss is formulated as Eq.15:

$$
L_{KL}^{KL} = -KL(q(\psi(z|I_c, h^1)||\mathcal{N}(0, 1))
$$

As for the generative path, $L_{KL}$ steer prior to close reconstructive path posterior, which can be formulated as Eq. 16:

$$
L_{KL}^{KL} = -KL(q(\psi(z|I_c, h^1)||p_\theta(z|h^0))
$$

Overall, the total loss function could be formulated as Eq. 17:

$$
L = \lambda_{KL}(L_{KL}^{KL} + L_{KL}^{KL}) + \lambda_1 L_1 + \lambda_T L_T
$$

(17)

where $L_{KL}$ regularizes KL divergence of prior and posterior distribution, $L_1$ maximize the expectation term from image quality and $L_T$ regularise text-image semantic similarity.

4 Experiments

4.1 Datasets

We evaluate our method on image captioning datasets CUB [Wah et al., 2011] and COCO [Lin et al., 2014], and use their original train, test and validation split. There are ten sentences per image in CUB and five in COCO. To further explore the inpainting quality of unique regions, we introduce object mask based on object boxes label on CUB. Completion of object mask regions is more challenging because of their uniqueness.

4.2 Implement Details

For images in both CUB and COCO datasets, we resize each training image to make its minimal length/width as 256 and crop the sub-image of size 256x256 at the center. All networks were initialized with orthogonal initialization [Saxe et al., 2013] and trained end-to-end with a learning rate of $10^{-4}$ and Adam optimizer [Kingma and Ba, 2014]. Match networks are pre-trained as in [Xu et al., 2018] on CUB and COCO respectively. During training, the weights of loss function terms are set as $\lambda_{KL} = \lambda_1 = 20$, $\lambda_T = 0.5$. Experiments are conducted on Ubuntu 18.04 system, with i7-9700K 3.70GHz CPU and 11G NVIDIA 2080Ti GPU.

We compare the proposed NIGDT model with two state-of-the-art methods PICNet [Zheng et al., 2019] and CSA [Liu et al., 2019] base on their official source code 1.

4.3 Quantitative Results

It has been mentioned [Yu et al., 2018; Yu et al., 2019] that image inpainting lacks good quantitative evaluation metrics. Still, we select commonly used mean $\ell_1$ loss, peak signal-to-noise ratio (PSNR), total variation (TV) and Structural Similarity (SSIM) for quantitative comparison. We will further discuss the semantic consistency of results in qualitative.

As the results in Table 1, the proposed NIGDT model outperforms state-of-the-art models on object mask in all measures, proving that this text-guided model works well in filling holes on such unique areas. On CUB and COCO dataset with the center mask, our model achieves comparable performance than these two inpainting models but does not exceed these models on two pixel-wise independence measures, which also explain our assumption because not all regions under the center mask are unique to surroundings and mentioned in the text. It is worth noting that the performance of all evaluated models decreases on object masked samples, while our model has the least performance degradation.

Table 1: Quantitative comparison with state-of-the-art on CUB and COCO dataset. ↓ means lower is better, ↑ means higher is better.

| Mask | Model | $\ell_1$ | PSNR↑ | TV loss↓ | SSIM↑ |
|------|-------|---------|--------|---------|--------|
| Center (CUB) | CSA | 7.99 | 20.95 | 9.51 | 0.824 |
| Center (COCO) | CSA | 7.73 | 21.57 | 9.79 | 0.824 |
| Object (CUB) | CSA | 15.85 | 19.13 | 8.33 | 0.689 |
| Ground Truth | 1.208 | 1.363 |

Table 2: Numerical result of user study.

| Model | Naturalness | Semantic Consistency |
|-------|-------------|----------------------|
| Ground Truth | 1.208 | 1.363 |
| CSA | 2.981 | 3.060 |
| PICNet | 3.178 | 3.469 |
| NIGDT | 2.531 | 2.106 |

Figure 4: Rank score distribution of different models.

1 https://github.com/KumapowerLIU/CSA-inpainting
https://github.com/lyndonzheng/Pluralistic-Inpainting
4.4 Qualitative Results

Fig. 3 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 photographic, but the CSA and PICNet fail to recover special characteristics in the original image, 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 surroundings such as color and texture. In comparison, the image generated by the NIGDT model recover these unique features as mentioned in the text and is 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.

4.5 User Study

To evaluate the models from human perspective, we conduct a user study through a ranking game. We collect 100 test images with captions with object masks from the CUB validation dataset. For each image, four solutions from (1) Ground truth, (2) Our model, (3) PICNet, (4) CSA model are prepared. We randomly shuffle the four samples and recruit 20 volunteers to rank these four images according to naturalness and the semantic consistency to the text description. On the four solutions, we compute the average of the ranking score. The results are shown in Table 2. According to the results, our model performs better than other models in terms of realistic, and significantly higher in semantic consistency.

4.6 Discussion

To investigate the effectiveness of our designed components including dual multi-modal attention, maxpooling on weighted word features, cross-modal match loss and text guidance, we trained four models without each component using the same super parameters and epochs. As shown in fig. 5, the model without guidance text provides an incomplete and semantically different solution. Without the maxpooling layer in dual multi-modal attention, the center of the generation area collapsed. When we remove match loss or dual multi-modal attention the models output still obtain the content of text but failed to allocate these characters to the exact location.

To further explore the usage of our work, we propose an interactive image editing case where users mask the region to be changed and give a text to describe a desired content. As in Figure 6, the top two examples are color variation edit by masking the belly or crown and give a text different from original image semantic. In the third case, we intentionally select a region between two differently colored areas and command the model to fill the hole as neighbor colors separately by two different sentences. The result shows the output qualify the text description and did not be affected by the color of adjacent areas. The experimental results presents a new possible formulation of image editing.

5 Conclusion

We proposed a neural image inpainting model guided with descriptive text (NIGDT). A dual multi-modal attention mechanism is integrated to the proposed model to exploit semantic features about the masked region from the descriptive text. What’s more, we presented an image-text matching based loss to improve the semantic consistency between inpainting output and the text. The experimental results on public datasets demonstrated that the proposed NIGDT model outperformed compared models, and generated significantly higher semantically consistent inpainting results. For future work, we will focus on inpainting more complex images with the guidance of texts.

References

[Bahdanau et al., 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine transla-
tion by jointly learning to align and translate. CoRR, abs/1409.0473, 2014.

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

[Bertalmio et al., 2002] Marcelo Bertalmio, Luminaita Vese, Sapiro T Guillermo, and Stanley Osher. Simultaneous structure and texture image inpainting. ITIP, 12(8), 2002.

[Doersch, 2016] Carl Doersch. Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908, 2016.

[Iizuka et al., 2017] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. TOG, 36:1–14, 07 2017.

[J. Weickert, 1999] J. Weickert. Coherence-enhancing diffusion filtering. volume 31, page 111–127, 1999.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv:1412.6980, 2014.

[Li et al., 2019a] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In Advances in Neural Information Processing Systems, pages 2063–2073, 2019.

[Li et al., 2019b] Bowen Li, Thomas Qi, Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. arXiv preprint arXiv:1912.06203, 2019.

[Lin et al., 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755. Springer, 2014.

[Liu et al., 2018] Guilin Liu, Fitsum Reda, Kevin Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In ECCV, page 85–100, 2018.

[Liu et al., 2019] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In ICCV, October 2019.

[Ma et al., 2019] Yuqing Ma, Xianglong Liu, Shihao Bai, Lei Wang, Dailan He, and Aishan Liu. Coarse-to-fine image inpainting via region-wise convolutions and non-local correlation. In IJCAI, pages 3123–3129, 2019.

[Miyato T, 2018] Koyama M et al Miyato T, Kataoka T. Spectral normalization for generative adversarial networks. 2018.

[Nazeri et al., 2019] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. 2019.

[Pathak et al., 2016] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. pages 2536–2544, 06 2016.

[S. Reed and Lee, 2016a] X. Yan L. Logeswaran B. Schiele S. Reed, Z. Akata and H. Lee. Generative adversarial text-to-image synthesis. 2016.

[S. Reed and Lee, 2016b] X. Yan L. Logeswaran B. Schiele S. Reed, Z. Akata and H. Lee. Generative adversarial text-to-image synthesis. 2016.

[Saxe et al., 2013] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. arXiv:1312.6120, 2013.

[Ting-Chun Wang and Catanzaro, 2018] Jun-Yan Zhu Andrew Tao Jan Kautz Ting-Chun Wang, Ming-Yu Liu and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. page 8798–8807, 2018.

[Vivek Kwatra and Kwatra, 2005] Aaron Bobick Vivek Kwatra, Irfan Essa and Nipun Kwatra. Texture optimization for example-based synthesis. volume 24, page 795–802, 2005.

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

[Xu et al., 2018] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. pages 1316–1324, 06 2018.

[Xudong Mao and Smolley, 2017] Haoran Xie Raymond YK Lau Zhen Wang Xudong Mao, Qing Li and Stephen Paul Smolley. Least squares generative adversarial networks. In CVPR, pages 2813–2821, 2017.

[Yeh et al., 2017] Raymond Yeh, Chen Chen, Teck Lim, Alexander Schwing, Mark Hasegawa-Johnson, and Minh Do. Semantic image inpainting with deep generative models. pages 6882–6890, 07 2017.

[Yu et al., 2018] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In CVPR, pages 5505–5514, 2018.

[Yu et al., 2019] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In CVPR, pages 4471–4480, 2019.

[Yuhang Song and Jay, 2018] Zhe Lin Xiaofeng Liu Qin Huang Hao Li Yuhang Song, Chao Yang and CC Jay. Contextual-based image inpainting: Infer, match, and translate. In ECCV, page 3–19, 2018.

[Zheng et al., 2019] Chuanxiao Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In CVPR, pages 1438–1447, 2019.