Unbiased Multi-Modality Guidance for Image Inpainting

Yongsheng Yu[1,3], Dawei Du[2], Libo Zhang[1,3,4⋆], and Tiejian Luo[3]

[1] Institute of Software Chinese Academy of Sciences, China
[2] Kitware, USA
[3] University of Chinese Academy of Sciences, China
[4] Nanjing Institute of Software Technology, China
yuyongsheng19@mails.ucas.ac.cn; cvdaviddo@gmail.com; libo@iscas.ac.cn;
tjluo@ucas.ac.cn

Abstract. Image inpainting is an ill-posed problem to recover missing or damaged image content based on incomplete images with masks. Previous works usually predict the auxiliary structures (*e.g.*, edges, segmentation and contours) to help fill visually realistic patches in a multi-stage fashion. However, imprecise auxiliary priors may yield biased inpainted results. Besides, it is time-consuming for some methods to be implemented by multiple stages of complex neural networks. To solve this issue, we develop an end-to-end multi-modality guided transformer network, including one inpainting branch and two auxiliary branches for semantic segmentation and edge textures. Within each transformer block, the proposed multi-scale spatial-aware attention module can learn the multi-modal structural features efficiently via auxiliary denormalization. Different from previous methods relying on direct guidance from biased priors, our method enriches semantically consistent context in an image based on discriminative interplay information from multiple modalities. Comprehensive experiments on several challenging image inpainting datasets show that our method achieves state-of-the-art performance to deal with various regular/irregular masks efficiently. The code is available at https://github.com/yeates/MMT.

Keywords: biased prior, multi-modality guidance, auxiliary denormalization, image inpainting

1 Introduction

Image inpainting aims to repair missing or damaged image content based on known information of an image. It has been applied on many real-world scenarios, such as image editing [1,3], unwanted object removal [8,30], and old photo restoration [31].

Following the assumption that corrupted images have adequate knowledge for inpainting [42,19], modern image inpainting methods [27,25,39,20,19] employ

[⋆] Corresponding author (libo@iscas.ac.cn).

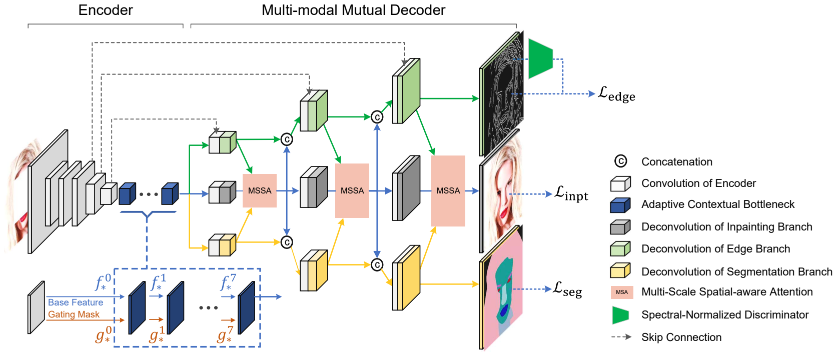
Encoder
Multi-modal Mutual Decoder
MSSA
MSSA
MSSA
c
$\mathcal{L}_{\text{edge}}$
$\mathcal{L}_{\text{inpt}}$
$\mathcal{L}_{\text{seg}}$
f_*^0
f_*^1
f_*^7
Base Feature
Gating Mask
g_*^0
g_*^1
g_*^7
Concatenation
Convolution of Encoder
Adaptive Contextual Bottleneck
Deconvolution of Inpainting Branch
Deconvolution of Edge Branch
Deconvolution of Segmentation Branch
MSA
Multi-Scale Spatial-aware Attention
Spectral-Normalized Discriminator
Skip Connection

Fig. 1. Architecture of our Multi-Modality guided Transformer that couples various modalities including RGB image, semantic segmentation, and edge textures.

an encoder-decoder architecture. Concretely, they focus on various contextual attention mechanisms to learn the known visible content and fill the missing region. However, this assumption does not hold if the image is damaged by larger masks. It is difficult to provide sufficient semantically consistent information for realistic image inpainting based on known area in a RGB image.

Therefore, recent approaches [39,25,32,21,4] have made great efforts to introduce auxiliary priors, such as *edges*, *segmentation*, and *contours*, to facilitate improving image inpainting performance. However, they still suffer from the biased prior issue by using predicted auxiliary structures to guide image inpainting intermediately. Without ground-truth in testing phase, such direct guidance is inevitably biased, resulting in more deviations and errors for image inpainting. On the other hand, previous works [32,23] are usually divided into multiple stages of neural networks under the U-Net architecture. If each stage contains a complex subnetwork, it is time-consuming for potential real-world inpainting applications. This problem becomes more prominent when extending to video inpainting. For example, Liu *et al.* [23] tackle the image inpainting problem by a two-stage process, *i.e.*, two individual U-Nets for rough inpainting and refinement inpainting, yielding the running speed of only 1.37 FPS.

To solve the above issues, we propose a new multi-modality guided transformer network for image inpainting. As shown in Fig. 1, it follows the U-Net style [28] encoder-decoder architecture. In the encoder, we first develop the adaptive contextual bottlenecks for better context reasoning. To adapt to the current image content and missing region, the gating mask is updated to weight different dilated convolutions to enhance base features. Then, the multi-modal mutual decoder is proposed to decode the enhanced features into three modalities, *i.e.*, RGB image, and corresponding semantic segmentation and edge textures. It consists of one image inpainting branch and two auxiliary branches for semantic segmentation and edge textures. Unlike existing approaches based on direct guidance from predicted auxiliary structures, we focus on jointly learning the unbiased discriminative interplay information among the three branches. Specifically, the proposed multi-scale spatial-aware attention mechanism integrates multi-modal feature maps via auxiliary denormalization to reduce duplicated and noisy content for image inpainting. Supervised by ground-truth RGB images, semantic segmentation and edge maps, the whole network is trained in an end-to-end fashion efficiently. Note that segmentation and edge annotations can be provided by the off-the-shelf algorithms [6,25].

As shown in Fig. 2, previous image inpainting methods fail to restore correct faces and buildings based on either biased edge [25] or segmentation [32] prior. On the contrary, our method still achieves robust results even though the glasses are not repaired in edge prior (see Ours* in the 1st row of Fig. 2) or the roof shape is not predicted correctly in segmentation prior (see Ours* in the 2nd row of Fig. 2). It demonstrates that our method can extract discriminative unbiased context information to guide image inpainting. To verify the effectiveness of our method, the experiment is conducted on three datasets including CelebA-HQ [17,18], OST [35] and CityScapes [7]. The results show our method achieves the state-of-the-art image inpainting performance. For example, our method obtains the best FID score on the CelebA-HQ dataset with both regular and irregular masks, yeilding ~ 2 gain over the second best performer CTSDG [12]. By using segmentation results from DeepLabv3+ [6], our method still performs well on those datasets without segmentation annotation (*e.g.*, Places2 [50]).

Contributions. 1. We propose an end-to-end multi-modality guided transformer to learn interplay information from multiple modalities including RGB image, edge textures and semantic segmentation. 2. We develop the multi-scale spatial-aware attention mechanism with auxiliary denormalization to capture compact and discriminative multi-modal features to guide unbiased image inpainting. 3. Comprehensive results on several datasets demonstrate the effectiveness of our unbiased multi-modality guidance, especially for irregular masks.

2 Related Work

Image inpainting. Mainstream image inpainting methods employ the encoder-decoder architecture based on the U-Net [28]. For example, Pathak *et al.* [27] introduces an adversarial network [11] to help train the U-Net and mitigate the blurring caused by the pixel-level averaging property of a reconstruction loss. After that, Contextual Attention (CA) [42] is a two-stage coarse-to-fine model to weight known region as the reference of mission region. Using partial conv [22], Recurrent Feature Reasoning (RFR) [19] applies multiple iterations at the bottleneck of the encoder from outside to inside for large corrupt areas. Different from partial conv [22] with a heuristic mask update step to standard convolution, Gated Conv (GC) [43] improves this mask update process with a learnable convolution layer.

To better exploit context between missing and uncorrupted regions, GLILC [16] first introduces multiple residual modules [13] of dilation convolution [41] as the bottleneck in the encoder. However, it may bring the “gridding” problem [5,34]

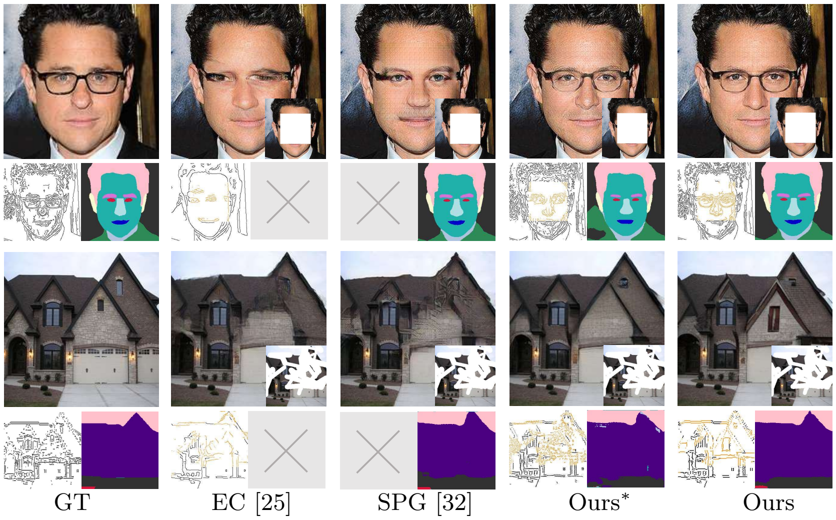

Fig. 2. Influence of biased prior guidance. × means no edge prior for SPG [32] and segmentation prior for EC [25]. Ours* denotes the variant of our multi-modality guided image inpainting method with inaccurate edge and segmentation priors by reducing the loss weights of two auxiliary branches by 30 times.

due to only sampling non-zero positions. That is, a single constant dilation rate results in either sparse convolution kernels (large hole rate) or difficulty crossing over large masks (small hole rate). To this end, Wang *et al.* [36] develop a generative multi-column network for image inpainting. Recently, Zeng *et al.* [46] propose the AOT blocks to aggregate contextual transformations from various receptive fields, which capture both informative distant image contexts and rich patterns of interest. Different from above methods, we introduce a new adaptive contextual bottleneck in the encoder, where the dynamic gating updating weights different pathways of dilated convolutions based on various masks.

Image inpainting with auxiliary structures. Due to the ill-posed nature of reconstructing missing regions, additional structural priors (*e.g.*, *edges*, *segmentation*, and *contours*) are used to facilitate image inpainting models for more realistic results. Edge Connect (EC) [25] relies on the corrupted canny edge image to deliver finer inpainting results. Cao and Fu [4] introduce an extra encoder to infer precise wireframe sketches to bypass the pool coherence of canny edge. According to the style and spatial consistency of semantic segmentation, Segmentation Prediction and Guidance network (SPG) [32] is a two-stage based segmentation and RGB image inpainting model, where DeepLabv3+ [6] is used to estimate the segmentation of corrupted image. Another work [39] is a new three-stage based model to locate and fill foreground object and its contour by disentangling the inter-object intersection.

However, the above multi-stage methods are usually time-consuming. For better efficiency, the Semantic Guidance and Evaluation (SGE) network [20] couples with segmentation and image inpainting at different layers of decoder, where the segmentation after completing and confidence scoring guides image inpainting by semantic normalization [26]. Liao *et al.* [21] propose the Semantic-wise Attention Propagation (SWAP) module to capture the semantic relevance between segmentation and image textures in non-local operation. Recently, Yang *et al.* [40] predict explicit edge embedding with an attention mechanism to facilitate image inpainting by the multi-task learning strategy. It worth mentioning that most aforementioned works use estimated auxiliary structures as the direct guidance of image inpainting. On the contrary, we develop the multi-head spatial-aware attention module to guide image inpainting based on jointly learned discriminative features from unbiased auxiliary priors.

Transformers in image inpainting. Inspired by Vision Transformer [10], recent methods [9,44] decode the long-range dependencies between input features for better image inpainting. Deng *et al.* [9] learn relations between the corrupted and uncorrupted regions and exploit their respective internal closeness. Yu *et al.* [44] introduce the bidirectional autoregressive transformer that enables bidirectionally modeling of contextual information of missing regions. In contrast, our method propose a new multi-modality guided transformer to capture interplay information across three modalities.

3 Multi-Modality Guided Transformer

The original image $\mathbf{I}$ is degraded as a corrupted image $\mathbf{I}_m = \mathbf{I} \odot (1 - \mathbf{M})$, where the pixel values in the missing region $\mathbf{M}$ equal to 0 are defined as invisible pixels. Our goal is to produce semantically reasonable and visually realistic reconstructed images $\mathbf{I}_{\text{pred}}$ with the input of the corrupted image $\mathbf{I}_m$. Similar to previous works [27,16,43,19], we retain the U-Net style encoder-decoder architecture. As illustrated in Fig. 1, the multi-modality guided transformer contains an encoder with adaptive contextual bottlenecks, and a multi-modal mutual encoder with multi-scale spatial-aware attention, described in detail as follows.

3.1 Encoder with Adaptive Contextual Bottlenecks

For better context reasoning, the multi-stream structure is used in the encoder to weight dilated convolutions and encode the current image content and missing region. Unlike simply stacking parameters in previous ASPP [5] and AOT [46], we develop a stack of Adaptive Contextual Bottlenecks (ACB) to adapt to the specific mask shape size and image context by dynamic gating. As shown in Fig. 3, the ACB module consists of four parallel pathways of convolutional layers with different dilation rate and one gating mask to weight dilated convolutions. In this way, the encoder can enlarge the perceptual field of convolutions and find the most plausible pathway according to the current missing region.

Given the corrupt image $\mathbf{I}_m$, the base features f_*^0 and gating g_*^0 are initialized by the last layer (gated conv) of encoder. Then f_*^l and g_*^l at each layer is updated

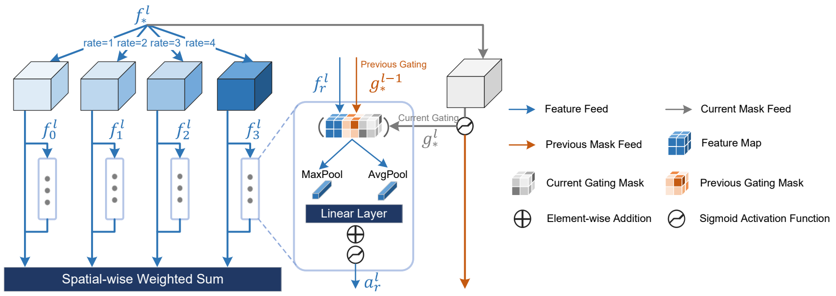

Fig. 3. Structure of Adaptive Contextual Bottlenecks in the encoder.

by the ACB block. The gating mask g_*^l is used to estimate the probability of missing region based on the feature map at the l-th layer ($l = 1, \cdots, L$), *i.e.*, $g_*^l = \text{gconv}(f_*^l)$, where gconv denotes the gated conv operation [43]. In terms of each pathway with dilation rate r, we compute the dilated feature maps f_r^l based on f_*^l and corresponding weight a_r^l. Similar to [38], the spatial-wise weight a_r^l is calculated based on both average and max pooling of concatenation of dilated feature maps f_r^l and gating masks g_*^l, g_*^{l-1}, *i.e.*, $a_r^l = \sigma(\text{fc}(\text{avg}(g_r^l)) + \text{fc}(\max(g_r^l)))$, where σ is the sigmoid function, and avg and max are the average and maximal pooling respectively. fc denotes the fully-connected layer, and the gating mask for each pathway is calculated as $g_r^l = \text{conv}([f_r^l; g_*^l; g_*^{l-1}])$. Finally, the feature map at the $(l+1)$-th ACB layer is updated by the spatial-wise weighted summation of f_r^l as

$$f_*^{l+1} = \sum_{r \in R} \frac{\exp(a_r^l)}{\sum_{r \in R} \exp(a_r^l)} \cdot f_r^l + f_*^l, \tag{1}$$

where R denotes the set of different dilation rates. The fractional term denotes element-wise product between dilated feature map f_r^l and attention vector a_r^l, weighting dilation block based on mask and image context. For simplicity, we omit the subscript l in the following sections.

3.2 Multi-modal Mutual Decoder

Given enhanced features f_*, the decoder use stacks of transformer blocks to learn the structural multi-modal information jointly. It consists of three branches, *i.e.*, one *inpainting branch* to recover the damaged image, and two *auxiliary branches* with additional segmentation and edge priors.

As shown in Fig. 1, within each transformer block, we first calculate the attention among feature maps from three branches by the proposed Multi-Scale Spatial-aware Attention (MSSA). Then, the enhanced features are split to combine the previous feature maps in each branch for attention calculation at next stage. Note that the skip connections between the encoder and decoder are used

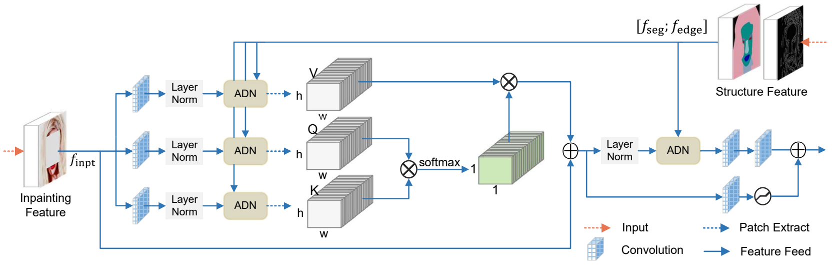

Fig. 4. Illustration of Multi-Scale Spatial-aware Attention.

to prevent network degradation. After three stages, we predict the inpainted image $\mathbf{I}_{\text{pred}}$, edge and segmentation maps. Thus we leverage the structural features from auxiliary branches to enforce the model focus on discriminative interplay features for more realistic image inpainting.

To learn mutual features from different modalities, it is intuitive to simply concatenate or add the feature maps in three branches. Nevertheless, such strategies may introduce duplicated and noisy content for image inpainting. To effectively integrate compact features from *auxiliary branches*, we introduce a new Multi-Scale Spatial-aware Attention (MSSA) mechanism as follows.

Multi-scale spatial-aware attention. Based on the encoded feature maps f_*, we use $f_{\text{inpt}}, f_{\text{edge}}, f_{\text{seg}}$ to denote the input feature maps for the inpainting branch, edge branch, and segmentation branch, respectively. As illustrated in Fig. 4, we combine the feature maps from three branches by the following Auxiliary DeNormalization (ADN):

$$\text{ADN}(f_{\text{inpt}}|[f_{\text{edge}}; f_{\text{seg}}]) = \gamma \odot \text{LN}(f_{\text{inpt}}) + \beta, \tag{2}$$

where $[;]$ denotes the matrix concatenation along channel dimension, and $\odot$ the element-wise multiplication. LN denotes layer normalization [2]. γ and β are the affine transformation parameters learned by two convolutional layers based on $[f_{\text{edge}}; f_{\text{seg}}]$ (see the top-right corner of Fig. 4). In this way, the multi-modal features are merged based on context from auxiliary structures that varies with respect to different spatial location.

Then, the merged features are embedded into query Q, key K and value V. Similar to [45], the embedded feature map is spatially split into N patches, *i.e.*, $P_i \in \mathbb{R}^{h\times w\times c} (i = 1, \ldots, N)$, where h, w, c denote the height, width and channel of patches respectively. The normalized self-attention $\alpha_{i,j}$ between patches i and j can be calculated as $\alpha_{i,j} = \text{softmax}(\frac{Q_i \cdot K_j^T}{\sqrt{h\cdot w\cdot c}}), \quad i, j \in 1, \ldots, N$. Note that we can perform multi-head self-attention like [10]. Thus the feature map of each patch is updated in a non-local form, *i.e.*, $\hat{P}_i = \sum_{j=1}^{N} \alpha_{i,j} V_j$.

Comparison between existing denormalization methods. Our ADN is related to two previous denormalization methods including AdaIN [15] and SPADE [26]. As shown in Fig. 5, we compare the networks of three denormalization methods. However, they are different in two aspects:

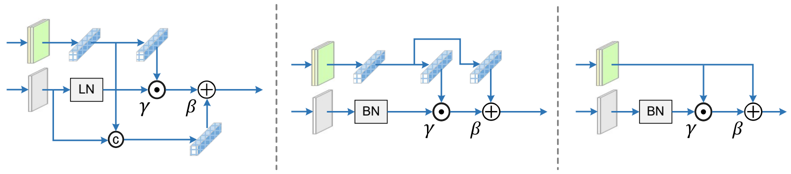

Fig. 5. (a) Our Auxiliary DeNormalization (ADN). (b) SPatially-Adaptive DEnormalization (SPADE) [26]. (c) Adaptive Instance Normalization (AaIN) [15]. LN, BN and IN denote layer, batch and instance normalizations respectively.

- AdaIN [15] and SPADE [26] learn the affine transformation parameters $\{\gamma, \beta\}$ based on the predicted auxiliary structures. Without ground-truth in testing phase, the predicted auxiliary structures are inevitably biased and result in inferior performance. In contrast, our ADN is based on the multi-modal features from two auxiliary branches.
- AdaIN [15] leverages the image's mean and variance instead of learnable affine parameters. SPADE [26] learns the spatial style of features by two convolutions after Batch Normalization. However, we combine features from both inpainting and auxiliary branches to learn the affine parameters.

Gated feed-forward. Finally, we piece all feature maps $\hat{P}_i$ together and reshape them with the original scale of input inpainting features f_{inpt}. Following the gated feed-forward layer, we can output the final feature maps for inpainted image prediction. Similar to gated conv [43], the gated feed-forward layer can ease the color discrepancy problem by detecting potentially corrupted and uncorrupted regions.

3.3 Optimization

To train our network, the overall loss consists of three terms, *i.e.*,

$$\mathcal{L} = \mathcal{L}_{\text{inpt}} + \lambda_{\text{edge}}\mathcal{L}_{\text{edge}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}, \tag{3}$$

where $\mathcal{L}_{\text{inpt}}$, $\mathcal{L}_{\text{edge}}$ and $\mathcal{L}_{\text{seg}}$ denote the loss terms for inpainting branch, edge branch and segmentation branch respectively. λ_{edge} and λ_{seg} are the balancing factors. The inpainting loss $\mathcal{L}_{\text{inpt}}$ follows the work in [22]. Similar to [25], we use both binary cross-entropy and adversarial loss functions to train the edge branch, *i.e.*,

$$\mathcal{L}_{\text{edge}} = w_1\mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{adv}}, \tag{4}$$

where w_1 is the balancing weight. $\mathcal{L}_{\text{BCE}} = \frac{1}{N}\sum_{i=1}^{N} -[\mathbf{C}_{\text{gt}}^i \log \mathbf{C}_{\text{pred}}^i + (1-\mathbf{C}_{\text{gt}}^i)\log(1-\mathbf{C}_{\text{pred}}^i)]$ predicts the edge structure, and $\mathcal{L}_{\text{adv}} = -\mathbb{E}\left[\mathbf{D}\left(\mathbf{C}_{\text{pred}}\right)\right]$ justifies if the predicted edge is fake or real. $\mathbf{C}_{\text{pred}}$ is the probability map between 0 and 1 for the reconstructed edge while $\mathbf{C}_{\text{gt}}$ is the ground-truth edge based on the canny operator [25]. $\mathbf{D}$ denotes the spectral normalization discriminator [24] that is composed of five convolutional layers. For the segmentation branch, we use the

Table 1. Quantitative comparison with the state-of-the-art approaches on CelebA-HQ. Easy, medium, and hard irregular masks denote the mask with coverage ratio of 10% $\sim$ 20%, 30% $\sim$ 40%, and 50% $\sim$ 60%, respectively. ↑ higher is better, and ↓ lower is better. Best and second best results are **highlighted** and $\underline{\text{underlined}}$.

mask type		irregular			regular
		easy	medium	hard	
PSNR↑	GC [43]	29.30	25.72	23.77	25.75
	RFR [19]	29.22	26.12	24.31	24.85
	CMGAN [48]	29.06	25.79	23.90	24.33
	ICT [33]	28.07	24.56	22.70	24.51
	CTSDG [12]	$\underline{29.59}$	$\underline{26.59}$	$\underline{24.69}$	$\underline{26.56}$
	Ours	**29.94**	**26.88**	**25.12**	**26.70**
SSIM↑	GC [43]	0.96	0.93	0.90	0.90
	RFR [19]	0.96	$\underline{0.94}$	0.91	0.87
	CMGAN [48]	**0.97**	$\underline{0.94}$	0.91	0.87
	ICT [33]	0.96	0.92	0.89	0.87
	CTSDG [12]	**0.97**	$\underline{0.94}$	$\underline{0.92}$	$\underline{0.91}$
	Ours	**0.97**	**0.95**	**0.93**	**0.92**
FID↓	GC [43]	15.00	18.41	21.28	22.45
	RFR [19]	7.37	10.74	13.45	14.35
	CMGAN [48]	6.80	11.85	14.12	12.91
	ICT [33]	$\underline{6.54}$	11.80	15.93	$\underline{11.90}$
	CTSDG [12]	7.80	$\underline{10.14}$	$\underline{13.30}$	14.52
	Ours	**6.47**	**9.32**	**11.61**	**11.40**

cross-entropy loss denoted by $\mathcal{L}_{\text{seg}} = \frac{1}{N}\sum_{i=1}^{N} -\mathbf{S}_{\text{gt}}^{i} \log \mathbf{S}_{\text{pred}}^{i}$, where $\mathbf{S}_{\text{gt}}^{i}$ and $\mathbf{S}_{\text{pred}}^{i}$ denote the ground-truth category and predicted probability for pixel i.

4 Experiment

We compare our method with state-of-the-arts on three large-scale datasets. An extensive ablation study is conducted to investigate the important designs in our model. All experiments are conducted on two 24G TITAN RTX GPUs.

Datasets. CelebA-HQ dataset [17,18] is a large-scale face image dataset with $30K$ HD face images, where each image has a semantic segmentation mask corresponding to 19 facial categories. Outdoor dataset (OST) [35] includes $9,900$ training images and 300 testing images for 8 semantic categories, which are obtained from the outdoor scene photography collection. Cityscapes dataset [7] contains $5,000$ street view images belonging to 20 categories. We expand the number of training images in this dataset, *i.e.*, $2,975$ images from the training set and $1,525$ images from the test set are used for training, and 500 images from the validation set are used for testing. In addition, the Places2 dataset [50] contains 10 million images covering more than 400 different types of scenes. We generate both regular and irregular masks to verify the ability of image inpainting methods. For regular masks, we draw a 128×128 centered square mask for CelebA-HQ and OST, and a 96×96 centered square mask for Cityscape. For irregular masks, we settle masks from [19] for CelebA-HQ and masks from [22] for Cityscape and OST.

Evaluation Metrics. Similar to the previous works [20,46], we use three metrics as follows. Peak Signal to Noise Ratio (PSNR) is an objective evaluation metric

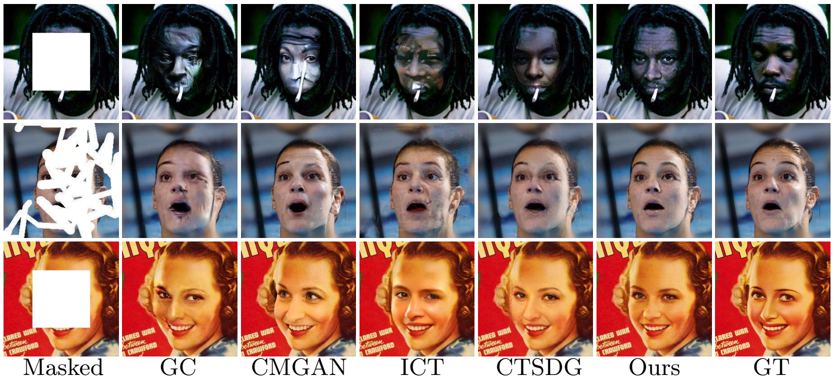

Fig. 6. Qualitative results of existing methods on CelebA-HQ.

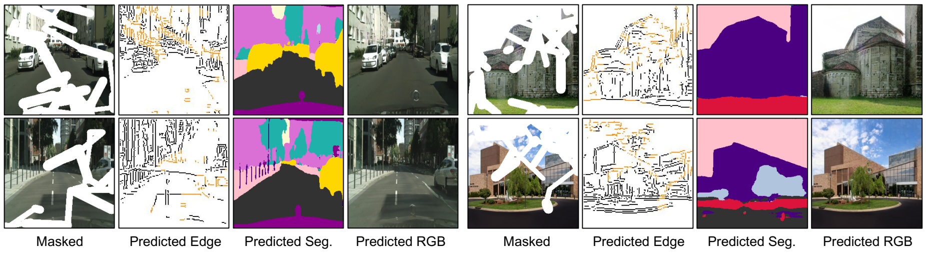

Fig. 7. Qualitative results of our method on Cityscape (1 to 4 columns) and OST (5 to 8 columns).

to assess the quality of generate images. Structural Similarity Index (SSIM) [37] uses the mean as an estimate of luminance, standard deviation as an estimate of contrast, and covariance as a measure of structural similarity to compare the difference between the generated and original images. Frechet Inception Distance (FID) [14] evaluates the accuracy and diversity of generated images. Notably, the Inception network [29] is used to extract the image features when calculating the FID score, and then calculate its mean and covariance matrix to estimate the distance between the ground-truth and generated data distribution. According to [47], deep metrics like FID are close to human perception.

4.1 Implementation Details

Our model is supervised by auxiliary structures including edge textures and semantic segmentation. With regard to edge structure, we employ the canny detection method [25] to generate edges of images. Besides, the CelebA-HQ, CityScapes and OST datasets all contain hand-crafted semantic segmentation, hence we can easily adopt these official labels for the segmentation part. More details of implementation are shown in the supplementary.

Table 2. Quantitative comparison with previous auxiliary prior guided approaches on OST and Cityscapes datasets.

method	auxiliary prior	OST						CityScapes					
		regular			irregular			regular			irregular		
		PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
EC [25]	edge	19.32	0.76	41.25	19.12	0.74	42.27	21.71	0.76	19.87	17.63	0.72	39.04
SPG [32]	seg.	18.04	0.70	45.31	17.85	0.74	50.03	20.14	0.71	23.21	16.41	0.67	43.63
SGINet [1]	seg.	-	-	-	-	-	-	25.74	0.87	23.02	18.53	0.77	57.53
SGE [20]	seg.	20.53	**0.81**	40.67	19.46	0.76	39.14	23.41	0.85	18.67	17.78	0.74	41.45
SWAP [21]	edge, seg.	21.18	**0.81**	**38.15**	20.31	0.80	36.74	23.89	0.84	18.14	17.86	0.76	38.18
Ours w/o seg.	edge	20.91	0.76	41.85	21.48	0.80	39.00	25.10	0.86	19.33	19.17	0.78	37.50
Ours w/o edge	seg.	21.80	0.77	40.96	22.58	0.81	36.03	25.95	0.87	17.85	**20.49**	**0.79**	34.79
Ours	edge, seg.	**21.84**	0.77	40.15	**23.15**	**0.82**	**35.77**	**26.13**	**0.88**	**17.52**	20.43	**0.79**	**33.45**

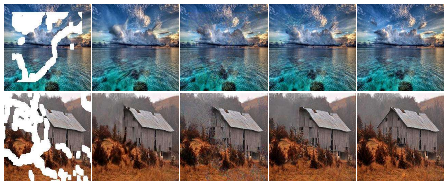

Fig. 8. Visual comparisons on Places2. From left to right: input, GC [43], EC [25], our method, and Ground Truth.

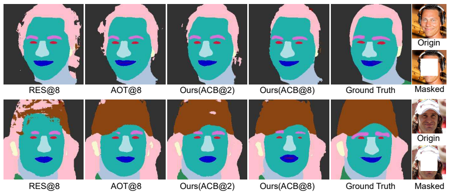

Fig. 9. Segmentation results with different bottlenecks on CelebA-HQ dataset with 128 × 128 regular center masks.

4.2 Result Analysis

We compare our model with several state-of-the-art methods including GC [43], RFR [19], CMGAN [48], ICT [33], CTSDG [12], SPG [32], SGINet [1], SGE [20], and SWAP [21]. A quantitative comparison is carried out on three datasets in terms of both regular and irregular masks with different coverage ratios. Full comparison results [46,23,49,25] we put in the appendix.

From Table 1, our method achieves the best or comparable performance among state-of-the-art image inpainting approaches that may not adopt auxiliary priors. Our method produces much better FID score than others for both regular and irregular masks, indicating that our inpainted results are more realistic. In Table 2, we compare several auxiliary prior guided inpainting approaches [25,32,20,21]. For a fair comparison with the methods relying on only one auxiliary structure, we construct two variants, denoted by "Ours w/o seg." and "Ours w/o edge". Compared with existing methods, our method achieves considerable gain respective to PSNR and FID especially on irregular masks. This is because our method focuses on the interplay representation from three modalities rather than directly guiding the image inpainting branch by predicted auxiliary structures (see Table 4).

In addition, we provide some visual examples on the CelebA-HQ dataset in Fig. 6. It can be seen that our method can generate more semantically consistent results compared with other approaches. More learned auxiliary priors of our method from CityScapes and OST datasets are visualized in Fig. 7.

Table 3. Contribution of two auxiliary branches in our method.

edge branch	segmentation branch	PSNR↑	SSIM↑	FID↓
✗	✗	25.88	0.90	12.36
✗	✓	26.47	0.91	11.42
✓	✗	26.19	0.90	11.95
✓	✓	**26.70**	**0.92**	**11.40**

Table 4. Comparison with different attention mechanisms.

variant	biased prior	attention mechanism	PSNR↑	SSIM↑	FID↓
MMT-1	✓	concat	26.17	0.89	20.01
MMT-2	✓	AdaIN [15]	26.17	0.89	21.71
MMT-3	✓	SPADE [26]	26.29	0.90	14.60
MMT-4	✓	MSSA+ADN	26.24	0.91	12.59
MMT-5	✗	MSSA+add	26.37	0.91	12.64
MMT-6	✗	MSSA+conv	26.50	0.91	11.90
MMT-7	✗	MSSA+AdaIN [15]	26.36	0.91	12.81
MMT-8	✗	MSSA+SPADE [26]	26.42	0.91	12.17
MMT-9	✗	MSSA+ADN	**26.70**	**0.92**	**11.40**

Additional results on Places2. Similar to SGE [20] and SWAP [21], we also conduct additional experiment on the Places2 dataset [50] for a comprehensive evaluation. Since there is no ground-truth segmentation, we use the segmentation results by DeepLabv3+ [6] to supervise the segmentation branch in our model. As shown in Fig. 8, the visual results show that our method still generate realistic inpainted images without ground-truth segmentation labels.

4.3 Ablation Study

To verify the effectiveness of the proposed modules in our network, the ablation experiments are carried out on the CelebA-HQ dataset.

Contribution of auxiliary branches. In Table 3, we construct three variants to verify the contribution of two auxiliary branches in our method. By learning from two auxiliary modalities, our method considerably outperforms the non-auxiliary variant w.r.t PSNR, SSIM, and FID. In addition, semantic segmentation contributes slightly more to image inpainting than edge textures. In summary, our Multi-Modal Mutual Decoder enriches semantic content on the inpainting branch by cross-attending segmentation and edge structures.

Biased prior guidance. Different from previous works [39,32,25,20,21] relying on biased prior guidance from predicted auxiliary structures, we jointly learn the interplay information of multi-modal features across the three branches and guide image inpainting based on ADN. To demonstrate its effectiveness, we construct four variants that are directly guided by predicted auxiliary structures.

In practice, we first add one convolutional layer at different stages to predict the auxiliary structures (Fig. 1), and then combine multi-modal features (Fig. 4).

In Table 4, MMT-1 denotes concatenating predicted structures with feature maps in the inpainting branch. MMT-2, MMT-3 and MMT-4 denote that we use AdaIN [15], SPADE [26], and MSSA with ADN to calculate the affine transformation parameters (γ, β) based on predicted structures, respectively. Compared with our method without biased prior guidance (*i.e.*, MMT-9), the FID score is significantly reduced based on predicted auxiliary structures. The results support our statement that predicted structures may introduce additional noises in image inpainting intermediately without ground-truth.

Effectiveness of multi-scale spatial-aware attention. To verify the effectiveness of Multi-Scale Spatial-aware Attention (MSSA), we construct four baseline feature fusion strategies from MMT-5 to MMT-8 in Table 4. MMT-5 means that we directly perform element-wise summation on features from three branches, while MMT-6 means that we splice the features from three branches together and then fuse them by two convolutional layers.

From Table 4, our MSSA performs the best in terms of three metrics. Compared with simple addition or convolution, our MSSA can provide reliable cross-attention among multiple modalities to guide high-quality reconstructed images. We also replace ADN by AdaIN [15] and SPADE [26] in MSSA for MMT-7 and MMT-8 respectively. The results show that our ADN performs better than previous normalization methods, demonstrating its effectiveness.

Effectiveness of adaptive contextual bottlenecks. In Table 5, we compare our Adaptive Contextual Bottlenecks (ACB) with the vanilla ResNet block [13] and the recently proposed AOT [46]. ACB@L ($L = 2, 4, 6, 8$) denotes L layers of ACB modules; RES@8 and AOT@8 denote 8 ResNet blocks [13] or 8 AOT blocks [46] respectively. † means quadrupling the channels of feature maps in ResNet blocks or copying base feature maps for different pathways in AOT blocks. The results show that the performance of ACB is improved along with the number of blocks is increased from 2 to 8. Using 8 ResNet or AOT blocks achieves similarly as that using 4 ACB blocks. It is worth mentioning that ResNet and AOT blocks have less number of channels of feature maps in each pathway. For a fair comparison, we construct two variants †RES@8 and †AOT@8 with the same channels as our ACB blocks. However, more channels in feature maps do not help improve the performance by using ResNet or AOT blocks. We speculate that the gating updating scheme in our ACB can reduce the influence of redundant noisy context with more channels of feature maps.

Besides, the mean of category-wise intersection-over-union (mIoU) [6] is another metric to validate the influence of bottleneck modules on segmentation inpainting. Our ACB module ($L \geq 4$) still outperforms other two blocks by more than 2%. The segmentation results in Fig. 9 also show that our ACB module generates more accurate segmentation performance. If the number of bottlenecks are increased, some isolated errors in segmentation can be removed (see the 3rd and 4th columns in Fig. 9).

Table 5. Comparison between different bottlenecks.

bottleneck	PSNR↑	SSIM↑	FID↓	mIoU%↑
RES@8	26.48	0.91	12.54	61.93
†RES@8	26.23	0.91	13.26	60.11
AOT@8	26.51	0.91	11.61	63.68
†AOT@8	26.29	0.91	14.17	62.28
ACB@2	26.48	0.91	12.18	63.54
ACB@4	26.60	0.91	12.24	65.84
ACB@6	26.61	0.91	12.09	66.16
ACB@8	**26.70**	**0.92**	**11.40**	**67.13**

Table 6. Efficiency of image inpainting networks.

method	params (M)	MACs (G)	speed (FPS)
SPG [32]	119.64	58.68	2.03
EC [25]	27.06	122.67	**67.21**
CTSDG [12]	52.15	**17.67**	36.99
RFR [19]	31.22	206.12	15.56
CSA [23]	132.11	55.23	1.37
RES@8 [13]	**22.76**	96.10	40.82
AOT@8 [46]	27.48	100.93	30.96
Ours (ACB@2)	**22.76**	96.10	40.88
Ours (ACB@8)	51.09	125.11	29.49

Efficiency comparison. From Table 6, we compare the number of parameters, computational complexity (MACs), and the running speed (FPS) of existing methods. Two-stage based SPG [32] and CSA [23], composed of complex sub-networks at each stage, run much more slowly than end-to-end methods. In contrast, EC [25] consists of two simple sub-networks for edge prediction and image inpainting, resulting in fast running speed but inferior performance. RFR [19] is an end-to-end model but predicts the inpainted results by the decoding heads recurrently. In terms of bottlenecks in the encoder, our ACB@2 achieves similar performance as AOT@8 with faster speed. By using 8 blocks, our method is still efficient with state-of-the-art performance among end-to-end methods.

Limitation discussion. Although our model generates promising results in most cases, it fails to recognize and recover unseen semantic knowledge, hence produces strange artifacts in complex scenes with large masks. Note that this weakness also affects other methods. It indicates that image inpainting model requires not only generative but also recognition capability. For example, our method can synthesize the human silhouette but lacks precise semantic details.

5 Conclusion

In this paper, we propose an end-to-end Multi-modality Guided Transformer for image impainting, which enriches coupled spatial features from shared multi-modal representations (*i.e.*, RGB image, semantic segmentation and edge textures). The proposed Multi-Scale Spatial-aware Attention can integrate compact discriminative features from multiple modalities via Auxiliary DeNormalization. Meanwhile, we introduce the Adaptive Contextual Bottlenecks in the encoder to enhance context reasoning for more semantically consistent inpainted results for the missing region. To the best of our knowledge, our scientific value lies in first analyzing the biased prior problem in image inpainting.

Acknowledgements and Declaration of Conflicting Interests. This work was supported by the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC038. Libo Zhang was supported Youth Innovation Promotion Association, CAS (2020111). Dr. Du and his employer received no financial support for the research, authorship, and/or publication of this article.

References

1. Ardino, P., Liu, Y., Ricci, E., Lepri, B., Nadai, M.D.: Semantic-guided inpainting network for complex urban scenes manipulation. In: ICPR. pp. 9280–9287 (2020)
2. Ba, L.J., Kiros, J.R., Hinton, G.E.: Layer normalization. CoRR **abs/1607.06450** (2016)
3. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: a randomized correspondence algorithm for structural image editing. TOG p. 24 (2009)
4. Cao, C., Fu, Y.: Learning a sketch tensor space for image inpainting of man-made scenes. In: ICCV (2021)
5. Chen, L., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. CoRR **abs/1706.05587** (2017)
6. Chen, L., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. pp. 833–851 (2018)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR. pp. 3213–3223 (2016)
8. Criminisi, A., Pérez, P., Toyama, K.: Object removal by exemplar-based inpainting. In: CVPR. pp. 721–728 (2003)
9. Deng, Y., Hui, S., Zhou, S., Meng, D., Wang, J.: Learning contextual transformer network for image inpainting. In: MM. pp. 2529–2538 (2021)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NeurIPS. pp. 2672–2680 (2014)
12. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: ICCV. pp. 14114–14123 (2021)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NeurIPS. pp. 6626–6637 (2017)
15. Huang, X., Belongie, S.J.: Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1510–1519 (2017)
16. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. TOG pp. 107:1–107:14 (2017)
17. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: ICLR (2018)
18. Lee, C., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR. pp. 5548–5557 (2020)
19. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: CVPR. pp. 7757–7765 (2020)
20. Liao, L., Xiao, J., Wang, Z., Lin, C., Satoh, S.: Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In: ECCV. pp. 683–700 (2020)
21. Liao, L., Xiao, J., Wang, Z., Lin, C., Satoh, S.: Image inpainting guided by coherence priors of semantics and textures. In: CVPR. pp. 6539–6548 (2021)

22. Liu, G., Reda, F.A., Shih, K.J., Wang, T., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: ECCV. pp. 89–105 (2018)
23. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: ICCV. pp. 4169–4178 (2019)
24. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: ICLR (2018)
25. Nazeri, K., Ng, E., Joseph, T., Qureshi, F.Z., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: ICCVW. pp. 3265–3274 (2019)
26. Park, T., Liu, M., Wang, T., Zhu, J.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR. pp. 2337–2346 (2019)
27. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: CVPR. pp. 2536–2544 (2016)
28. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)
29. Salimans, T., Goodfellow, I.J., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: NeurIPS. pp. 2226–2234 (2016)
30. Shetty, R., Fritz, M., Schiele, B.: Adversarial scene editing: Automatic object removal from weak supervision. In: NeurIPS. pp. 7717–7727 (2018)
31. Song, L., Cao, J., Song, L., Hu, Y., He, R.: Geometry-aware face completion and editing. In: AAAI. pp. 2506–2513 (2019)
32. Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., Kuo, C.J.: Spg-net: Segmentation prediction and guidance network for image inpainting. In: BMVC. p. 97 (2018)
33. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. In: ICCV. pp. 4672–4681 (2021)
34. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.W.: Understanding convolution for semantic segmentation. In: WACV. pp. 1451–1460 (2018)
35. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: CVPR. pp. 606–615 (2018)
36. Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.: Image inpainting via generative multi-column convolutional neural networks. In: NeurIPS. pp. 329–338 (2018)
37. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP pp. 600–612 (2004)
38. Woo, S., Park, J., Lee, J., Kweon, I.S.: CBAM: convolutional block attention module. In: ECCV. vol. 11211, pp. 3–19 (2018)
39. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., Luo, J.: Foreground-aware image inpainting. In: CVPR. pp. 5840–5848 (2019)
40. Yang, J., Qi, Z., Shi, Y.: Learning to incorporate structure knowledge for image inpainting. In: AAAI. pp. 12605–12612 (2020)
41. Yu, F., Koltun, V., Funkhouser, T.A.: Dilated residual networks. In: CVPR. pp. 636–644 (2017)
42. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: CVPR. pp. 5505–5514 (2018)
43. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: ICCV. pp. 4470–4479 (2019)
44. Yu, Y., Zhan, F., Wu, R., Pan, J., Cui, K., Lu, S., Ma, F., Xie, X., Miao, C.: Diverse image inpainting with bidirectional and autoregressive transformers. In: MM. pp. 69–78 (2021)
45. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: ECCV. pp. 528–543 (2020)

46. Zeng, Y., Fu, J., Chao, H., Guo, B.: Aggregated contextual transformations for high-resolution image inpainting. CoRR **abs/2104.01431** (2021)
47. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
48. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. In: ICLR (2021)
49. Zheng, C., Cham, T., Cai, J.: Pluralistic image completion. In: CVPR. pp. 1438–1447 (2019)
50. Zhou, B., Lapedriza, À., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. TPAMI pp. 1452–1464 (2018)