Learning a Sketch Tensor Space for Image Inpainting of Man-made Scenes

Chenjie Cao, Yanwei Fu
School of Data Science, Fudan University
{20110980001,yanweifu}@fudan.edu.cn

(a) Inputs (b) Exemplar-based (c) GateConv (d) EdgeConnect (e) Ours

Figure 1: Results of (b) Exemplar-based inpainting [2], (c) GateConv [43], (d) EdgeConnect [28], and (e) our method, with the inpainted areas highlighted by red dot box. Our method makes better inpainting without artifacts.

Abstract

This paper studies the task of inpainting man-made scenes. It is very challenging due to the difficulty in preserving the visual patterns of images, such as edges, lines, and junctions. Especially, most previous works are failed to restore the object/building structures for images of man-made scenes. To this end, this paper proposes learning a Sketch Tensor (ST) space for inpainting man-made scenes. Such a space is learned to restore the edges, lines, and junctions in images, and thus makes reliable predictions of the holistic image structures. To facilitate the structure refinement, we propose a Multi-scale Sketch Tensor inpainting (MST) network, with a novel encoder-decoder structure. The encoder extracts lines and edges from the input images to project them into an ST space. From this space, the decoder is learned to restore the input images. Extensive experiments validate the efficacy of our model. Furthermore, our model can also achieve competitive performance in inpainting general nature images over the competitors.

1. Introduction

As a long-standing problem, image inpainting has been studied to address the problem of filling in the missing parts of the images being semantically consistent and visually realistic with plausible results. Thus, image inpainting is useful to many real-world applications, *e.g.* image restoration, image editing, and object removal [6].

Intrinsically as an inverse problem, the inpainting is challenging in both restoring the missed global structure (*semantically consistent*), and generating realistic regions locally coherent to unmasked regions (*visually consistent*). Especially, it is hard to reconstruct the missed image regions from the complex man-made scenes and structures, due to the difficulty in preserving the prominent low-level visual patterns, such as edges, line segments, and junctions, as shown in Fig. 1. To this end, this paper particularly focuses on learning to reconstruct these visual patterns for image inpainting, and proposes a method of the best merits in repairing the masked regions of man-made scene images, such as images with indoor and outdoor buildings.

Both traditional approaches [2, 8, 19] and deep learning methods [24, 44, 40, 17, 43, 21] had made great efforts on reconstructing the structures of images in producing visually realistic results. However, these methods are still challenged by producing structurally coherent results, especially in the inpainting of man-made scenes. Typically, inpainting approaches may suffer from the following problems. (1) *Missing critical structures.* Traditional synthesis-based approaches are normally unable to model the critical structures as in Fig. 1(b). On the other hand, recent learning-based inpainting methods utilize auxiliary information to support the inpainting, *e.g.*, edges [28, 20], and segmentation [32, 22], predominantly inpainting local visual cues, rather than holistic structures of man-made scenes. For example, the results of EdgeConnect [28] with canny edges [5] in Fig. 1(d) suffer from broken and blurry line segments and lose connectivity of building structures. (2) *Unreliable pattern transfer.* The learning-based auxiliary detectors will transfer and magnify the unreliable image priors or patterns to the masked image regions, which causes degraded inpainting results [22]. (3) *Trading off performance and efficiency.* The auxiliary-based inpainting methods usually consume more computation, due to additional components or training stages [28, 20, 39]. But these methods still have artifacts in junction regions as framed with red dotted lines in Fig. 1(c)(d). Therefore, the design of a more effective network is expected to efficiently enhance the inpainting performance.

To address these issues, our key idea is to learn a Sketch Tensor (ST) space by an encoder-decoder model. (1) The encoder learns to infer both local and holistic critical structures of input images, including canny edges and compositional lines. The image is encoded as the binarized 'sketch' style feature maps, dubbed as *sketch tensor*. The decoder takes the restored structure to fill in holes of images. (2) For the first time, the idea of parsing wireframes [14] is re-purposed to facilitate inpainting by strengthening the holistic structures of man-made scenes with more effective and flexible holistic structures. We propose a Line Segment Masking (LSM) algorithm to effectively train the wireframe parser, which alleviates unreliable structure guidance from corrupted images and the heavy computation of auxiliary detectors during the training phase. Besides, LSM also leverage the separability of line segments to extend the proposed model to obtain better object removal results. (3) Most importantly, we significantly boost the training and inference process of previous inpainting architectures. Thus, a series of efficient modules are proposed, which include partially gated convolutions, efficient attention module, and Pyramid Decomposing Separable (PDS) blocks. Critically, we present PDS blocks to help better learn binary line and edge maps. Our proposed modules make a good balance of model performance and training efficiency.

Formally, this paper proposes a novel Multi-scale Sketch Tensor inpainting (MST) network with an encoder-decoder structure. The encoder employs LSM algorithm to train an hourglass wireframe parser [38] and a canny detector to extract line and edge maps. These maps concatenated with input images are projected into ST space by Pyramid Structure Sub-encoder (PSS), which is sequentially constructed by 3 partially gated convolution layers, 8 dilated residual block layers with an efficient attention module, and 3 pyramid decomposing block layers as in Fig. 2. The image is encoded as a third-order sketch tensor in ST space, representing local and holistic structures. Finally, the decoder is stacked by two groups of 3 partially gated convolution layers for both ends with 8 residual block layers, which will re-project the sketch tensor into the restored image.

We highlight several contributions here. (1) We propose learning a novel sketch tensor space for inpainting tasks. Such a space is learned to to restore the critical missed structures and visual patterns, and makes reliable predictions of the holistic image structures. Essentially, the skech tensor has good interpretability of input images, as empirically shown in experiments. (2) For the first time, the wireframe parsing has been re-purposed to extract lines and junctions for inpainting. A novel line segment masking algorithm is proposed to facilitate training our inpainting wireframe parser. (3) We introduce the novel partially gated convolution and efficient attention module to significantly improve model performance without incurring additional expensive computational cost than competitors, such as EdgeConnect. (4) A novel pyramid decomposing separable block is proposed to address the issue of effectively learning sparse binary edge and line maps. (5) Extensive experiments on the dataset of both man-made and natural scenes, including ShanghaiTech [14], Places2 [48], and York Urban [4]. show the efficacy of our MST-net, over the competitors.

2. Related work

Image Inpainting by Auxiliaries. The auxiliary information of semantics and structures have been utilized to help inpainting tasks in traditional methods, such as lines, structures [13, 1], and approximation images [10]. Recently, deep learning based approaches take auxiliary information as the important prior, such as canny edges [28, 20], smoothed edges [29, 12], edge and gradient information [39], and semantic segmentation [32, 22]. Auxiliary information has been also utilized in image editing [47, 30, 15, 18]. Particularly, EdgeConnect [28] and GateConv [43] both leverage edges to inpaint masked areas with specific structural results. However, there is no holistic structure information for man-made scenes that has been explicitly modeled and utilized as auxiliary information in previous work. Such information is crucial to inpaint the images of complex structures, such as buildings, or indoor furniture. To this end, our method utilizes the separability and connectivity of line segments to help improve the performance of inpainting.

Line Detection Approaches. Line detection enjoys immense value in many real-world problems. Therefore, it has been widely researched and developed in computer vision. Many classic graphics algorithms are proposed to extract line segments from raw images [9, 34, 26]. However, these traditional methods suffer from intermittent and spu-

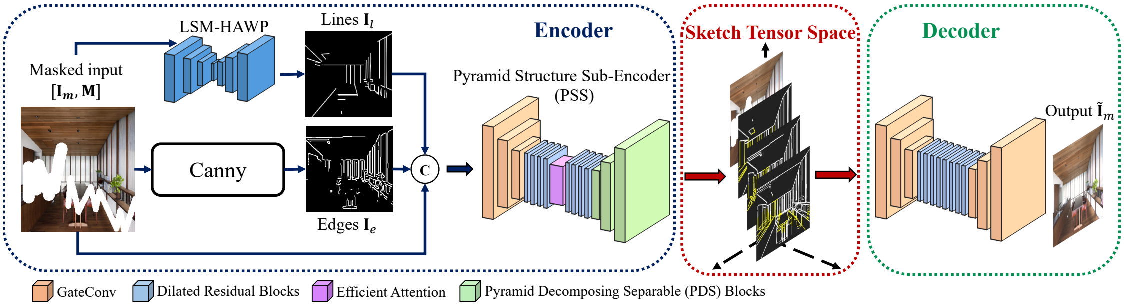

Figure 2: The overview of MST, which is consisted of encoder, sketch tensor space, and decoder.

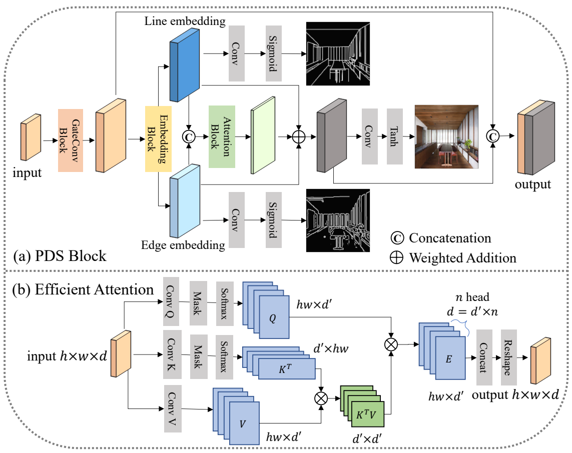

Figure 3: The illustration of (a): Pyramid Decomposing Separable (PDS) block, (b): Efficient Attention block.

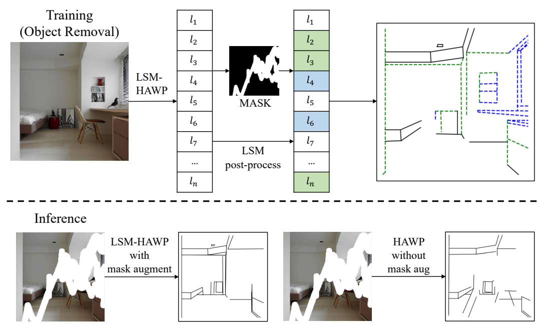

Figure 4: LSM Illustration. In training, lines segments are denoted by Eq. (1). Black solid lines, blue and green dotted lines indicate the retained, masked, and masked by probably m lines. Only in inpainting training process or object removal task, the ground-truth images are known in advance.

rious line segments. Moreover, the extracted lines lack positions of junctions which causes poor connectivity. Huang *et al*. [14] propose a deep learning based wireframe parsing to improve the line detection task, which uses DNNs to predict heatmaps of junctions and line segments as the wireframe representation. LCNN proposed in [49] leverages heuristic sampling strategies and a line verification network to improve the performance of wireframe parsing. Furthermore, Xue *et al*. [38] utilize the holistic attraction field to enhance both efficiency and accuracy for the wireframe parser. Although wireframe parsing based models enjoy distinctive strengths, to the best of our knowledge, no previous work has considiered utilizing wireframe parser to help preserve line structures for downstreaming inpainting task. Critically, for the first time, our method first repurposes the wireframe parser to facilitate image inpainting, which leverages connectivity and structural information of wireframes to achieve better inpainting performance.

3. Multi-scale Sketch Tensor Inpainting

Overview. The MST network is shown in Fig. 2. Given the input masked image $\mathbf{I}_m \in \mathbb{R}^{h\times w\times 3}$ and corresponding binary mask $\mathbf{M}$, MST has three key components, *i.e.*, encoder $\mathbf{\Phi} : [\mathbf{I}_m, \mathbf{M}] \to \mathcal{S}$, decoder $\mathbf{\Psi} : \mathcal{S} \to \tilde{\mathbf{I}}_m$, and Sketch Tensor (ST) space of a third-order tensor denoted by $\mathcal{S} \in \mathbb{R}^{h\times w\times 3}$. Particularly, the encoder firstly employs the improved wireframe parser LSM-HAWP and canny detector [5] to extract line $\mathbf{I}_l$ and edge maps $\mathbf{I}_e$; then the concatenated image and maps $[\mathbf{I}_l; \mathbf{I}_e; \mathbf{I}_m; \mathbf{M}]$ is processed by Pyramid Structure Sub-Encoder (PSS) to produce the ST space $\mathcal{S}$. The decoder predicts inpainted image $\tilde{\mathbf{I}}_m$, closer to ground-truth image $\mathbf{I}$.

In this section, Sec. 3.1 will introduce the LSM algorithm. Details about the PSS and ST space are specified in Sec. 3.2. Finally, the decoder will be discussed in Sec. 3.4.

3.1. Line Segment Masking Algorithm

The wireframe parser HAWP [38] is adopted to extract lines from images. Specifically, it extracts the junction set $\mathbf{J} = \left\{\mathbf{p} = (x, y) \in \mathbb{R}^2\right\}$, and line set $\mathbf{L} = \{\mathbf{l} = (\mathbf{p}_a, \mathbf{p}_b) = (x_a, y_a, x_b, y_b)\}$ paired by junctions in $\mathbf{J}$. Unfortunately, if lines are corrupted by the mask $\mathbf{M}$, the results of naive HAWP [38] will be largely degraded as shown in the inference stage of Fig. 4 with broken structures.

To this end, we propose an LSM algorithm to infer missed line segments with the flexible wireframe parsing, which is composed of two parts. 1) Learning an LSM-HAWP network by retraining HAWP with the irregular [43] and object segmentation [45] masks. This not only improves the stability for the corrupted images, but also achieves superior results in trivial wireframe parsing tasks

(Sec. 4.2). 2) Introducing an indicator function to denote the masking probability of each line segment $\mathbf{l}$ according to the mask $\mathbf{M}$ as the post-processing for LSM-HAWP,

$$\mathbb{I}(\mathbf{l}) = \begin{cases} 1 & if\ \mathbf{p}_a \in \mathbf{M}\ and\ \mathbf{p}_b \in \mathbf{M} \\ 0 & if\ \mathbf{p}_a \notin \mathbf{M}\ and\ \mathbf{p}_b \notin \mathbf{M}\ , \\ m & otherwise \end{cases} \quad (1)$$

where m $(0 \leq m \leq 1)$ is a hyper-parameter between 0 and 1, and junction $\mathbf{p} \in M, \mathbf{p} \notin M$ means that $\mathbf{p}$ is masked or not. Fig. 4 illustrates the LSM post-process in training and inference stages. In training, we set $m = 0.5$ to train the first half epoches, and $m = 1$ for training in the rest epoches. In the object removal task, we use the unmasked image with $m = 0$ to retain necessary structures.

Thus as one practical strategy of speedup the learning process, we extract all wireframes beforehand and filtering them by the post-process of LSM, rather than an end-to-end training with LSM-HAWP, which dramatically improves the training efficiency. Note that in the testing stage for inpainting, corrupted images are used as the input to make our results fair comparable to the competitors.

3.2. Pyramid Structure Sub-Encoder

The LSM-HAWP and canny detector extract the line and edge maps $\mathbf{I}_l, \mathbf{I}_e \in \mathbb{R}^{h\times w\times 1}$ respectively. Essentially, $\mathbf{I}_e$ is binary map, and $\mathbf{I}_l$ is got from connecting junction pairs from LSM-HAWP with anti-aliased lines. The input to PSS is the concatenation of the masked image and structure maps $[\mathbf{I}_l; \mathbf{I}_e; \mathbf{I}_m]$. As shown in Fig. 2, the PSS is composed of partially gated convolutions, dilated [41] residual blocks, efficient attention, and pyramid decomposing separable blocks, which will be explained next. For the detailed structures, please refer to our supplementary.

Partially Gated Convolutions. We adopt the Gated Convolution (GC) layers to process the masked input features, as it works well for the irregular mask inpainting tasks in [43]. Unfortunately, GC demands much more trainable parameters than vanilla convolutions. To this end, as shown in Fig. 2 and Fig. 3(a), we propose a partially GC strategy of only utilizing three GC layers for the input and output features in both encoder and decoder models. Essentially, this is motivated by our finding that the outputs of GC mostly devoting to filtering features of masked and unmasked regions only in the encoder layers of the coarse network and the decoder layers of the refinement network in [43]. In contrast, we do not observe significantly performance improvement of using GC in the middle layers of backbones. Thus, we maintain vanilla convolutions (i.e., residual blocks) in the middle layers, to save parameters and improve the performance as empirically validated in Tab. 4.

Efficient Attention Block. Intuitively, attention is important to learn patterns crossing spatial locations in image inpainting [42, 25, 21, 39]. However, attention modules are expensive to be computed, and non-trivial to be parallelized in image inpainting [42]. To this end, we leverage Efficient Attention (EA) module among middle blocks of Fig. 2 and detailed in Fig. 3(b). Particularly, for the input feature $\mathbf{X} \in \mathbb{R}^{h\times w\times d}$, and mask $\mathbf{M} \in \mathbb{R}^{h\times w\times 1}$, we first reshape them to $\mathbb{R}^{hw\times d}$ and $\mathbb{R}^{hw\times 1}$ as the vector data respectively. Then, the process of efficient attention can be written as

$$\begin{aligned} \mathbf{Q} &= \mathrm{softmax}_{row}(\mathbf{W}_q\mathbf{X} + (\mathbf{M}\cdot -\infty)) \\ \mathbf{K} &= \mathrm{softmax}_{col}(\mathbf{W}_k\mathbf{X} + (\mathbf{M}\cdot -\infty)) \\ \mathbf{V} &= \mathbf{W}_v\mathbf{X}, \qquad \mathbf{E} = \mathbf{Q}(\mathbf{K}^T\mathbf{V}), \end{aligned} \quad (2)$$

where $\mathbf{W}_{q,k,v}$ indicates different learned parameter matrices; $+(\mathbf{M}\cdot -\infty)$ means masking the corrupted inputs before the softmax operation. Critically, $\mathrm{softmax}_{row}$ and $\mathrm{softmax}_{col}$ means the softmax operations on the row and column individually. Then, we achieve the output $\mathbf{E} \in \mathbb{R}^{hw\times d}$, which will be reshaped back to $\mathbb{R}^{h\times w\times d}$. Note that Eq. (2) is an approximation to vanilla attention operation as in [31]. Typically, this strategy should in principle, reduce significantly computational cost. Since $\mathbf{E}$ is computed by $\mathbf{K}^T\mathbf{V} \in \mathbb{R}^{d\times d}$, rather than standard $\mathbf{Q}\mathbf{K}^T \in \mathbb{R}^{hw\times hw}$. In practice, the dimension d is much smaller than hw in computer vision tasks. Furthermore, as in Fig. 3(b), we introduce the multi-head attention [33] to further reduce the dimension from d to $d' = d/n_{head}$. Thus, $\mathbf{K}^T\mathbf{V}$ could aggregate the query $\mathbf{Q}$ in feature level and obtain the global context $\mathbf{E}$. Note that EA module is inspired but different from [31], as the attention scores of corrupted regions are masked to aggregate features from uncorrupted regions.

Pyramid Decomposing Separable (PDS) Block. We have found that GAN-based edge inpainting [35] suffers from generating meaningless edges or unreasonable blanks for masked areas as shown in Fig. 6, due to the nature of sparsity in binary edge maps. Thus, we propose a novel PDS block in PSS as in Fig. 3(a). PDS leverages dilated lower-scale structures to make up the sparse problem for ST space. Particularly, assume the image feature $\mathbf{X} \in \mathbb{R}^{h\times w\times d}$, PDS firstly decouples it as

$$\begin{aligned} &\mathbf{E}_{le} = f_{eb}(\mathrm{GateConv}(\mathbf{X})),\ \{\mathbf{E}_l, \mathbf{E}_e\} = \mathrm{split}(\mathbf{E}_{le}), \\ &\mathbf{O}_e = \sigma(\mathrm{Conv}(\mathbf{E}_e)),\ \mathbf{O}_l = \sigma(\mathrm{Conv}(\mathbf{E}_l)), \end{aligned} \quad (3)$$

where σ is sigmoid activation. We project the feature into two separated embedding spaces called line embedding $\mathbf{E}_l$ and edge embedding $\mathbf{E}_e$ with the function f_{eb}, which is composed with $\mathrm{Conv2D} \to \mathrm{IN} \to \mathrm{ReLU}$. split is the operation of splitting the features into two respective tensors with equal channels. Then, $\mathbf{E}_l$ and $\mathbf{E}_e$ are utilized to learn to roughly reconstruct the image $\mathbf{O}_{im}$ without corruption as

$$\begin{aligned} \mathbf{A} &= f_{ab}(\mathbf{E}_{le}) \in \mathbb{R}^{h\times w\times 1} \\ \mathbf{E}'_{le} &= \mathbf{E}_l \odot (1 - \mathbf{A}) + \mathbf{E}_e \odot \mathbf{A} \\ \mathbf{O}_{im} &= \tanh(\mathrm{Conv}(\mathbf{E}'_{le})), \end{aligned} \quad (4)$$

where f_{ab} is $\mathrm{Conv2D} \to \mathrm{IN} \to \mathrm{ReLU} \to \mathrm{Conv2D} \to \mathrm{Sigmoid}$. $\odot$ denotes the element-wise product. The prediction of coarse $\mathbf{O}_{im}$ gives a stronger constraint to the conflict of $\mathbf{E}_l$ and $\mathbf{E}_e$, and we expect the attention map $\mathbf{A}$ can be learned adaptively according to a challenging target. Furthermore, a multi-scale strategy is introduced in the PDS. For the given feature $\mathbf{X}^{(1)} \in \mathbb{R}^{64\times 64\times d}$ from dilated residual blocks of PSS, PDS predicts lines, edges, and coarse images with different resolutions as follow

$$\mathbf{X}^{(i+1)}, \mathbf{O}_{im}^{(i)}, \mathbf{O}_{l}^{(i)}, \mathbf{O}_{e}^{(i)} = \mathrm{PDS}^{(i)}(\mathbf{X}^{(i)}), \tag{5}$$

where $i = 1, 2, 3$ indicate that the output maps with 64×64, 128×128, and 256×256 resolutions respectively. Various scales of images can be in favor of reconstructing different types of structure information.

Loss Function of PSS. We minimize the objectives of PPS with two spectral norm [27] based SN-PathGAN [43] discriminators D_l and D_e for lines and edges respectively. And the $\mathcal{L}_D^{enc}$ and $\mathcal{L}_G^{enc}$ are indicated as

$$\begin{aligned} \mathcal{L}_D^{enc} =& \mathcal{L}_{D_l} + \mathcal{L}_{D_e}, \\ \mathcal{L}_G^{enc} =& \lambda_a \mathcal{L}_{adv}^{enc} + \lambda_f \mathcal{L}_{fm} + \sum_{i=1}^{3} \|\mathbf{O}_{im}^{(i)} - \mathbf{I}^{(i)}\|_1, \end{aligned} \tag{6}$$

where $\mathbf{O}_{im}^{(i)}$, $\mathbf{I}^{(i)}$ are the multi-scale outputs (Eq. 5) and ground-truth images varying image size 64×64 to 256×256 respectively. $\mathcal{L}_{fm}$ is the feature matching loss as in [28] to restrict the l_1 loss between discriminator features from the concatenated coarse-to-fine real and fake sketches discussed below. And the adversarial loss can be further specified as

$$\begin{aligned} \mathcal{L}_{adv}^{enc} &= -\mathbb{E}\left[\log D_l(\hat{\mathbf{O}}_l)\right] - \mathbb{E}\left[\log D_e(\hat{\mathbf{O}}_e)\right], \\ \mathcal{L}_{D_l} &= -\mathbb{E}\left[\log D_l(\hat{\mathbf{I}}_l)\right] - \mathbb{E}\left[1 - \log D_l(\hat{\mathbf{O}}_l)\right], \\ \mathcal{L}_{D_e} &= -\mathbb{E}\left[\log D_e(\hat{\mathbf{I}}_e)\right] - \mathbb{E}\left[1 - \log D_e(\hat{\mathbf{O}}_e)\right], \end{aligned} \tag{7}$$

where $\hat{\mathbf{O}}_l$, $\hat{\mathbf{O}}_e \in \mathbb{R}^{256\times 256\times 3}$ are got from the multi-scale PSS outputs of lines and edges, which are upsampled and concatenated with $\hat{\mathbf{O}}_l = [\mathrm{up}(\mathbf{O}_l^{(1)}); \mathrm{up}(\mathbf{O}_l^{(2)}); \mathbf{O}_l^{(3)}]$ and $\hat{\mathbf{O}}_e = [\mathrm{up}(\mathbf{O}_e^{(1)}); \mathrm{up}(\mathbf{O}_e^{(2)}); \mathbf{O}_e^{(3)}]$. To unify these outputs into the same scale, the nearest upsampling $\mathrm{up}()$ is utilized here. Accordingly, $\hat{\mathbf{I}}_l$, and $\hat{\mathbf{I}}_e$ are the concatenation of multi-scale ground-truth edge and line maps, respectively. Note that, we firstly dilate the ground-truth edges and lines with the 2×2 kernel, and then subsample them to produce the low resolution maps at the scale of 64×64, 128×128. The hyper-parameters are set as $\lambda_a = 0.1$, $\lambda_f = 10$.

3.3. Sketch Tensor (ST) Space

The last outputs $\mathbf{O}_l^{(3)}, \mathbf{O}_e^{(3)} \in \mathbb{R}^{256\times 256\times 1}$ of Eq. 5 is used to compose the ST space as

$$\mathcal{S} = \left[\mathbf{O}_l^{(3)}; \mathbf{O}_e^{(3)}; \mathrm{clip}(\mathbf{O}_l^{(3)} + \mathbf{O}_e^{(3)})\right]. \tag{8}$$

Generally, lines represent holistic structures, while edges indicate some local details. They provide priors of structures in different manners to the inpainting model. We also combine and clip them within 0 and 1 to emphasize overlaps and make an intuitive expression of the whole structure.

3.4. Decoder of MST Network

The structure of the decoder is the same as PSS except that the decoder has no attention blocks and PSD blocks. Because we find that the generator in [28] is sufficient to generate fine results with reasonable structure information. But we still add gated mechanism to the leading and trailing three convolutions to improve the performance for the irregular mask. As the encoder has provided the ST space $\mathcal{S} \in \mathbb{R}^{h\times w\times 3}$ in Eq. 8, we can get the inpainted image $\tilde{\mathbf{I}}_m$ with the masked input $\mathbf{I}_m$ and mask $\mathbf{M}$ as

$$\tilde{\mathbf{I}}_m = \mathbf{\Psi}([\mathbf{I}_m; \mathbf{M}; \mathcal{S}]). \tag{9}$$

We use SN-PathGAN [43] based discriminator D_{im} for the decoder, and the objectives to be minimized are

$$\begin{aligned} \mathcal{L}_D^{dec} &= -\mathbb{E}\left[\log D_{im}(\mathbf{I})\right] - \mathbb{E}\left[1 - \log D_{im}(\tilde{\mathbf{I}}_m)\right], \\ \mathcal{L}_G^{dec} &= \mathcal{L}_{l_1} + \lambda_a \mathcal{L}_{adv}^{dec} + \lambda_p \mathcal{L}_{per} + \lambda_s \mathcal{L}_{style}, \end{aligned} \tag{10}$$

$$\mathcal{L}_{adv}^{dec} = -\mathbb{E}\left[\log D_{im}(\tilde{\mathbf{I}}_m)\right], \tag{11}$$

where $\mathbf{I}$ is the origin undamanged image for training. We adopt the l_1 loss $\mathcal{L}_{l_1}$, VGG-19 based perceptual loss [16] $\mathcal{L}_{per}$, and style-loss [7] $\mathcal{L}_{style}$ to train our model, with the empirically setted hyperparameters $\lambda_p = 0.1$, $\lambda_s = 250$ [28, 12]. We propose using the balanced loss form [3] to implement reconstruction losses $\mathcal{L}_{l_1}$ and $\mathcal{L}_{per}$, normalizing by the weights of masked region $\mathbf{M}$ and unmasked region $1 - \mathbf{M}$. The balanced loss form can settle the inpainting task with imbalanced irregular masks properly.

Note that we learn the encoder and the decoder jointly in the forward pass. The gradients of the decoder will not be propagated to update the encoder, which not only saves the memory but also maintains the interpretability of the sketch tensor $\mathcal{S}$. Generally, the total parameter number and calculation costs of our model are comparable to [28].

4. Experiments and Results

Datasets. The proposed approach is evaluated on three datasets: ShanghaiTech [14], Places2 [48], and York Urban [4]. ShanghaiTech contains 5000 training images and 462 test images consisted of buildings and indoor scenes with wireframe labels. The LSM-HAWP wireframe detection model is trained on this dataset with the mask augmentation. For Places2, we select 10 categories images according to the number of line segments detected by HAWP as man-made scenes (P2M). Moreover, we randomly select 10 Places2 classes with both natural and urban scenes as comprehensive scenes (P2C)[1]. For York Urban, it has 102 city street view images for testing models trained with P2M.

[1] Details about Places2 are illustrated in the supplementary.

Ground Truth | Input | GC | EC | RFR | MED | Ours | Ours*

Figure 5: Qualitative results in ShanghaiTech and man-made Places2, where * means that our method works in the LMS object removal mode with $m = 0$. Key parts are enlarged, and the complete generated pictures are shown in the supplementary.

		GC	EC	RFR	MED	Ours	Ours*
S.-T.	P.↑	25.54	26.51	26.30	26.14	**26.90**	27.20
	S.↑	**0.893**	0.871	0.863	0.886	0.876	0.880
	F.↓	26.06	24.84	25.95	30.02	**21.16**	20.05
P2M	P.↑	25.72	26.71	26.07	23.84	**27.20**	27.47
	S.↑	0.892	0.901	0.890	0.857	**0.907**	0.910
	F.↓	16.54	14.75	17.79	26.95	**12.81**	12.13
Y.-U.	P.↑	25.92	26.13	25.64	24.20	**26.29**	26.59
	S.↑	**0.886**	0.864	0.852	0.858	0.869	0.872
	F.↓	31.68	32.06	38.70	51.71	**29.15**	26.53
P2C	P.↑	27.87	28.35	–	–	**28.52**	28.65
	S.↑	0.923	0.927	–	–	**0.928**	0.929
	F.↓	15.05	13.68	–	–	**11.97**	11.69

Table 1: The PSNR (P.), SSIM (S.) and FID (F.) results on ShanghaiTech (S.-T.), man-made Places2 (P2M), York Urban (Y.-U.), and comprehensive Places2 (P2C). ↑ means larger is better and ↓ means lower is better. * indicates our method working in object removal mode of LSM. Best results are bold except for object removal ones.

Implementation Details. Our method is implemented with PyTorch in 256×256 image size. All models are trained with Adam optimizer of $\beta_1 = 0$ and $\beta_2 = 0.9$, and the initial learning rates are $2e-4$ and $2e-5$ for generators and discriminators respectively. We train the model with 400k steps in ShanghaiTech and 1000k steps in Places2. Besides, E_s is trained with 1/3 of the total steps, and then be fixed. We decay the learning rate with 0.75 for each 100k steps. For the structure information, line segments are extracted by HAWP, and Canny edges are got with $\sigma = 2$. Our model is trained in Pytorch v1.3.1, and costs about 2 days training in ShanghaiTech and about 5 days in Places2 with a single NVIDIA(R) Tesla(R) V100 16GB GPU.

Comparison Methods. We compare the proposed MST with some state-of-the-art methods, which include Gated Convolution (GC) [43], Edge Connect (EC) [28], Recurrent Feature Reasoning (RFR) [21], and Mutual Encoder-Decoder with Feature Equalizations (MED) [12]. These methods are all retrained with similar settings and costs compared with ours.

Settings of Masks. To handle the real-world image inpainting and editing tasks, such as object removal, the random irregular mask generation in [43, 37] is adopted in this paper. Besides, as discussed in [45], real-world inpainting tasks usually remove regions with typical objects or scenes segments. So we collect 91707 diverse semantic segmentation masks with various objects and scenes with the coverage rates in $[5\%, 40\%]$ from the COCO dataset [23]. Overall, the final mask will be chosen from irregular masks and COCO segment masks randomly with 50% in both training and test set. To be fair, all comparison methods are retrained with the same masking strategy.

4.1. Image Inpainting Results

For fair comparisons in image inpainting, we do not leak any line segments of the uncorrupted images for the image inpainting task and related discussions. The wireframe parser LSM-HAWP is trained in ShanghaiTech, and it predicts the line segments for the other two datasets. Besides, results from the object removal with $m = 0$ in Eq. (1) are offered for reference only in this section.

Quantitative Comparisons. In this section, we evaluate results with PSNR, SSIM [36], and FID [11]. As discussed

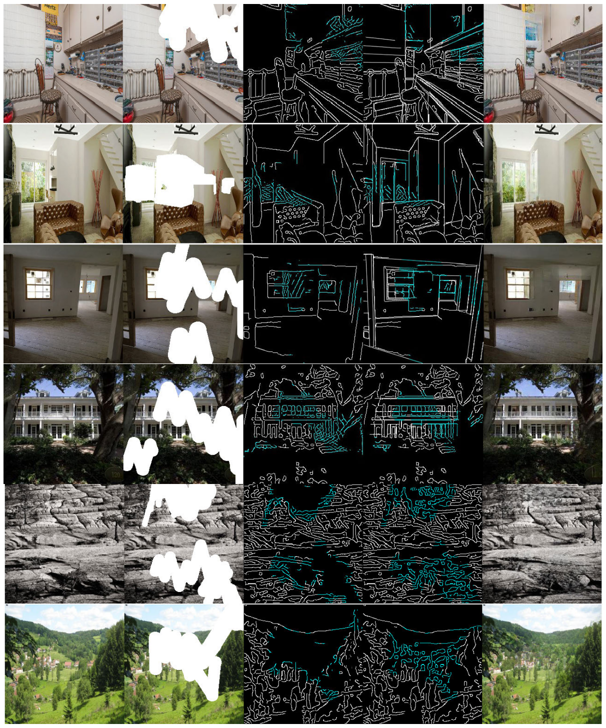

Figure 6: (From left to right) Original image, masked image, edges $\mathbf{O}_e$ generated from EC [28], edges and lines $\text{clip}(\mathbf{O}_l + \mathbf{O}_e)$ generated from our model, our inpainted results. Blue lines indicate the generated lines or edges.

in [46], we find that l_2 based metrics such as PSNR and SSIM often contradict human judgment. For example, some meaningless blurring areas will cause large perceptual deviations but small l_2 loss [46]. Therefore, we pay more attention to the perceptual metric FID in quantitative comparisons. The results on ShanghaiTech, P2M, and York Urban are shown in Tab. 1. From the quantitative results, our method outperforms other approaches in PSNR and FID. Especially for the FID, which accords with the human perception, our method achieves considerable advantages. Besides, our method enjoys good generalization, as it also works properly in P2M, York Urban, and even P2C.

Qualitative Comparisons. Qualitative results among our method and other state-of-the-art methods are shown in Fig. 5. Compared with other methods, our approach achieves more semantically coherent results. From the enlarged regions, our schema significantly outperforms other approaches in preserving the perspectivity and the structures of man-made scenes. Moreover, we show the edge results of EC [28] and partial sketch tensor space (edges and lines) results of our method in ShanghaiTech, P2M, and P2C in Fig. 6. Results in Fig. 6 demonstrate that line segments can supply compositional information to avoid indecisive and meaningless generations in edges. Furthermore, line segments can also provide clear and specific structures

	GC	EC	Ours
S.-T.	3.00	7.67	**33.67**
P2M	6.67	8.67	**32.33**

Table 2: Average user scores of ShanghaiTech (S.-T.) and man-made Places2 (P2M).

in masked regions to reconstruct more definitive results in man-made scenes. Besides, for the natural scene images in P2C without any lines, our method can still outperform EC with reasonable generated edges for the masked regions. As discussed in Sec. 3.2, the proposed PDS works properly to handle the sparse generative problem.

Results of Natural Scenes. In the last three rows of Tab. 1, we present the results of ours method on the comprehensive Places2 dataset (P2C) to confirm the generalization. They are compared with GC and EC, which have achieved fine scores in man-made Places2 (P2M). Note that all metrics are improved in P2C compared with ones in P2M of Tab. 1, which demonstrates man-made scenes are more difficult to tackle. Our methods still get the best results among all competitors, and the object removal results achieve superior performance. Besides, the last two rows of Fig. 6 show that our method can generate reliable edge sketches even without lines in natural scenes. These phenomenons are largely due to two reasons: 1) There is still a considerable quantity of line segments in the comprehensive scenes, and these instances are usually more difficult than others. 2) The proposed partially gated convolutions, efficient attention, and PDS blocks can work properly for various scenes.

Human Judgements. For more comprehensive comparisons, 50 inpainted images from GC [43], EC [28], and ours are chosen from ShanghaiTech and P2M randomly. And these samples are compared by 3 uncorrelated volunteers. Particularly, volunteers need to choose the best one from the mixed pool of the inpainted images with different methods, and give one score to the respective approach. If all methods work roughly the same for a certain sample, it will be ignored. The average scores are shown in Tab. 2. Ours method is significantly better than other competitors.

4.2. Other Applications and Ablation Study

Object Removal. From the last column in Fig. 5, we show the inpainting results of our MST working in the post-process of LSM with $m = 0$ for the object removal. These refined images with more reasonable structures indicate that the our model can strengthen the capability of image reconstruction without redundant residues. From Fig. 7, some object removal cases are shown to confirm the practicability of the proposed method. Numerical results are presented in the last columns of Tab. 1. So, our method can achieve further improvements in all metrics compared with the vanilla image inpainting schema for the object removal, and it can correctly mask undesired lines. Therefore, the post-process

Figure 7: Object removal examples. From left to right: origin image, masked image, and our inpainted image.

	unmasked testset			masked testset		
Threshold	5	10	15	5	10	15
HAWP	62.16	65.94	67.64	35.39	38.47	40.15
LSM-HAWP	**63.20**	**67.06**	**68.70**	**48.93**	**53.30**	**55.39**

Table 3: Structural average precision scores [49] with different thresholds of parsers trained with (LSM-HAWP) and without (HAWP) the mask augmentation on ShanghaiTech.

	PSNR↑	SSIM↑	FID↓	Param
w/o GC	26.51	0.871	24.84	10.8M
Partially GC	**26.78**	**0.874**	22.68	12.1M
All GC	26.71	**0.874**	**22.62**	25.1M

Table 4: Ablation study of the gated convolution of our model in ShanghaiTech, Param indicates the parameter scale of the compared inpainting model.

	w/o lines	w/o EA	w/o PDS	Ours
PSNR↑	26.78	26.77	26.63	**26.90**
SSIM↑	0.874	0.875	0.873	**0.876**
FID↓	22.68	21.39	21.88	**21.16**

Table 5: Quantitative ablation studies in ShanghaiTech.

of LSM can significantly improve the performance in the object removal task with the separability of line segments .

Masked Wireframe Detection. As discussed in Sec. 3.1, we retrain the HAWP with mask augmentation to ensure the robustness for the corrupted data as LSM-HAWP in ShanghaiTech dataset [14]. To confirm the effect of the augmentation, we further prepare another masked ShanghaiTech testset with the same images. The results are shown in Tab. 3. The structural average precision (sAP) metric is proposed in [49], which is defined as the area under the PR curve of detected lines with different thresholds. From Tab. 3, LSM-HAWP performs better than vanilla HAWP significantly for the masked testset. Moreover, HAWP can also gain improvements in the uncorropted testset with more than 1% in sAP, which indicates the great generality of the mask augmentation for the wireframe detection task.

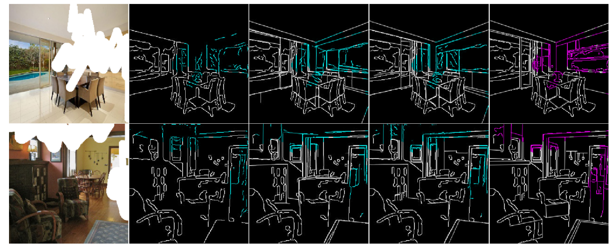

Figure 8: Ablation studies on line segments and PDS blocks in ShanghaiTech. Blue and red lines indicate recovered and ground truth in masked regions, respectively.

Simplification of Gated Convolutions. The related ablation study of GC are shown in Tab. 4. Only replacing vanilla convolutions for input and output stages of the model with partially GC gains satisfactory improvements. But replacing all convolution layers with GC fails to achieve a further large advance, while the parameters are doubled[2].

Contributions of Lines, PDS, and EA. We explored the effects of lines, Pyramid Decomposing Separable (PDS) blocks and the Efficient Attention (EA) module in Tab. 5 in ShanghaiTech. Specifically, line segments, PDS blocks and the EA module are removed respectively, while other settings unchanged. As shown in Fig. 8, our method without line segments causes severe structural loss, and the one without PDS suffers from contradictory perspectivity. From Tab. 5, we can see all line segments, PDS blocks and the EA module can improve the performance in image inpainting.

5. Conclutions

This paper studies an encoder-decoder MST-net for inpainting. It learns a sketch tensor space restoring edges, lines, and junctions. Specifically, the encoder extracts and refines line and edge structures into a sketch tensor space. The decoder recover the image from it. Moreover, the LSM algorithm is proposed to update the line extractor HAWP to fit the inpainting and object removal tasks. Several effective network modules are proposed to improve the MST for tough man-made scenes inpainting. Extensive experiments validate the efficacy of our MST-net for inpainting.

Acknowledgment Yanwei Fu is the corresponding authour. This project is supported by Huawei, and released on https://ewrfcas.github.io/MST_inpainting.

[2]The visualization results of GC are shown in the supplementary.

References

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 2

[2] Antonio Criminisi, Patrick Perez, and Kentaro Toyama. Object removal by exemplar-based inpainting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–II. IEEE, 2003. 1, 1

[3] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019. 3.4

[4] Patrick Denis, James H Elder, and Francisco J Estrada. Efficient edge-based methods for estimating manhattan frames in urban imagery. In *European conference on computer vision*, pages 197–210. Springer, 2008. 1, 4

[5] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern Recognition*, 34(3):721–725, 2001. 1, 3

[6] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, pages 1–22, 2019. 1

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 3.4

[8] Qiang Guo, Shanshan Gao, Xiaofeng Zhang, Yilong Yin, and Caiming Zhang. Patch-based image inpainting via two-stage low rank approximation. *IEEE transactions on visualization and computer graphics*, 24(6):2023–2036, 2017. 1

[9] Peter E Hart and RO Duda. Use of the hough transformation to detect lines and curves in pictures. *Communications of the ACM*, 15(1):11–15, 1972. 2

[10] James Hays and Alexei A Efros. Scene completion using millions of photographs. *Communications of the ACM*, 51(10):87–94, 2008. 2

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 4.1

[12] Bin Jiang Hongyu Liu, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, 2020. 2, 3.4, 4

[13] Jia-Bin Huang, Sing Bing Kang, Narendra Ahuja, and Johannes Kopf. Image completion using planar structure guidance. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014. 2

[14] Kun Huang, Yifan Wang, Zihan Zhou, Tianjiao Ding, Shenghua Gao, and Yi Ma. Learning to parse wireframes in images of man-made environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 626–635, 2018. 1, 2, 4, 4.2

[15] Youngjoo Jo and Jongyoul Park. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1745–1753, 2019. 2

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 3.4

[17] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas. Prior guided gan based semantic inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[18] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 2

[19] Haodong Li, Weiqi Luo, and Jiwu Huang. Localization of diffusion-based inpainting in digital images. *IEEE Transactions on Information Forensics and Security*, 12(12):3050–3064, 2017. 1

[20] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5962–5971, 2019. 1, 2

[21] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768, 2020. 1, 3.2, 4

[22] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shinichi Satoh. Guidance and evaluation: Semantic-aware image inpainting for mixed scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 683–700. Springer, 2020. 1, 2

[23] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 4

[24] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 1

[25] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4170–4179, 2019. 3.2

[26] Xiaohu Lu, Jian Yao, Kai Li, and Li Li. Cannylines: A parameter-free line segment detector. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 507–511. IEEE, 2015. 2

[27] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative ad-

versarial networks. In *International Conference on Learning Representations*, 2018. 3.2

[28] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 1, 1, 2, 3.2, 3.4, 3.4, 4, 6, 4.1, 4.1

[29] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 181–190, 2019. 2

[30] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5400–5409, 2017. 2

[31] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient attention: Attention with linear complexities. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3531–3539, 2021. 3.2

[32] Yuhang Song, Chao Yang, Yeji Shen, Peng Wang, Qin Huang, and C-C Jay Kuo. Spg-net: Segmentation prediction and guidance network for image inpainting. *arXiv preprint arXiv:1805.03356*, 2018. 1, 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3.2

[34] Rafael Grompone Von Gioi, Jeremie Jakubowicz, Jean-Michel Morel, and Gregory Randall. Lsd: A fast line segment detector with a false detection control. *IEEE transactions on pattern analysis and machine intelligence*, 32(4):722–732, 2008. 2

[35] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3.2

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4.1

[37] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5840–5848, 2019. 4

[38] Nan Xue, Tianfu Wu, Song Bai, Fudong Wang, Gui-Song Xia, Liangpei Zhang, and Philip HS Torr. Holistically-attracted wireframe parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2788–2797, 2020. 1, 2, 3.1

[39] Jie Yang, Zhiquan Qi, and Yong Shi. Learning to incorporate structure knowledge for image inpainting. In *AAAI*, pages 12605–12612, 2020. 1, 2, 3.2

[40] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[41] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3.2

[42] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 3.2

[43] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 1, 1, 2, 3.1, 3.2, 3.2, 3.4, 4, 4.1

[44] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1486–1494, 2019. 1

[45] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 3.1, 4

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4.1

[47] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. Real-time user-guided image colorization with learned deep priors. *ACM Transactions on Graphics (TOG)*, 9(4), 2017. 2

[48] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1, 4

[49] Yichao Zhou, Haozhi Qi, and Yi Ma. End-to-end wireframe parsing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 962–971, 2019. 2, 3, 4.2