Hourglass Attention Network for Image Inpainting

Ye Deng[1], Siqi Hui[1], Rongye Meng[1], Sanping Zhou[1,2], and Jinjun Wang[1]

[1] Xi'an Jiaotong University, Xi'an, China
[2] Shunan Academy of Artificial Intelligence, Ningbo, China
{dengye, huisiqi}@stu.xjtu.edu.cn
mengrongye@gmail.com
spzhou@xjtu.edu.cn
jinjun@mail.xjtu.edu.cn

Abstract. Benefiting from the powerful ability of convolutional neural networks (CNNs) to learn semantic information and texture patterns of images, learning-based image inpainting methods have made noticeable breakthroughs over the years. However, certain inherent defects (e.g. local prior, spatially sharing parameters) of CNNs limit their performance when encountering broken images mixed with invalid information. Compared to convolution, attention has a lower inductive bias, and the output is highly correlated with the input, making it more suitable for processing images with various breakage. Inspired by this, in this paper we propose a novel attention-based network (transformer), called hourglass attention network (HAN) for image inpainting, which builds an hourglass-shaped attention structure to generate appropriate features for complemented images. In addition, we design a novel attention called Laplace attention, which introduces a Laplace distance prior for the vanilla multi-head attention, allowing the feature matching process to consider not only the similarity of features themselves, but also distance between features. With the synergy of hourglass attention structure and Laplace attention, our HAN is able to make full use of hierarchical features to mine effective information for broken images. Experiments on several benchmark datasets demonstrate superior performance by our proposed approach. The code can be found at github.com/dengyecode/hourglassattention.

Keywords: image inpainting, attention, transformer

1 Introduction

Image inpainting [3] is the process of filling missing areas of an image with reasonable content. It can support many applications such as removing objects, restoring old photos, image editing, etc. For image inpainting, it is most critical to be able to give plausible content to fill the target region based on the observed region and make the whole image consistent.

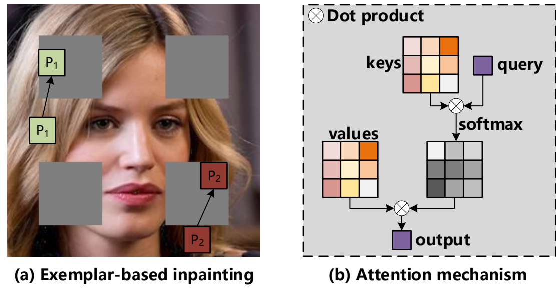

Fig. 1. Illustration of the connection between exemplar-based methods and attention mechanisms. Exemplar-based methods often try to find the appropriate content for broken regions in visible regions based on certain prior, while the result of the attention is obtained by weighting the "value" based on the similarity between the "query" and the "key". Thus we can consider the attention mechanism as a special exemplar-base method, where the "key-value" pairs play the role of exemplar.

Traditional exemplar-based methods [1,26,2,21] usually match and copy background patches into missing areas or by propagating information from boundaries around the missing regions. These methods are quite effective for images with only a small portion of breakage or repeated patterns, while they often fail to generate reasonable results for images with large broken regions or complex structures due to the lack of higher-level semantic understanding of the image.

In recent years, benefiting from the advantages of convolution neural networks (CNNs) for representation learning, learning-based approaches [39,22,58,28] have made noticeable breakthroughs. Nonetheless, CNNs have some limitations in complementing broken images. Firstly, each filter of CNNs spatially shares convolution kernel parameters when dealing with the broken input. For a single image with both broken and normal areas, each vanilla convolution operator allocates identical kernels for both valid, invalid as well as mixed (e.g. the ones located on broken border) features (pixels), which easily leads to structural distortions, texture blurring and artifacts, especially when the patterns are complex or the damaged regions are vast [30,59]. Secondly, CNNs that operate only within a local window are inefficient at modeling the long-range structure of an image, while in the processing of image inpainting, proper information within the entire image, sometimes far away from the corrupted regions, needs to be utilized for corrupted regions.

To relieve the above limitation, we propose to learn an Hourglass Attention Network (HAN) for image inpainting, which builds an hourglass-shaped attention structure based on the powerful texture pattern learning capability of CNNs to mine the contextual information in the hierarchical features to generate appropriate feature maps for the reconstructed images. Compared to convolution, the attention module has a lower inductive bias and is able to generate different weights depending on the miscellaneous input, thus making it more suitable and flexible for images with multiple breakages in the inpainting tasks. Besides, there is a close connection between the attention based and the exemplar-based methods in terms of borrowing information from within the image. Specifically, exemplar-based approaches try to find the most plausible content to fill the target (unknown) areas based on the observed region of the image. As for the (dot-product) attention, the result is based on the relationship between the "query" and the "key-value" pairs, and we can consider the "key-value" as a special kind of exemplar, as shown in Figure 1.

Our Hourglass Attention Network (HAN) consists of three parts, including a CNN encoder, a CNN decoder, and the hourglass attention structure. Particularly, the encoder is a stack of multiple convolution layers. It can be considered as a learnable feature extractor, which is responsible for the input images into the feature maps. The decoder, on the other hand, is similar in structure to the encoder and corresponds to the task of decoding the feature map into output images. As for the hourglass attention structure, it consists of attention blocks designed for feature sequences of different patch sizes stacked in a certain order, based on the property that a feature map can be divided into sequences of different patch sizes. To be specific, we divide the hourglass attention structure into two parts, the feature encoding and the feature decoding. In the feature encoding part, we employ the attention blocks from small to large according to the patch size, while in the feature decoding part we place the blocks from large to small. Therefore, the feature map resolution (number of patches) decreases gradually in the feature encoding phase and increases gradually in the feature decoding phase, which is similar to autoencoder or U-net [44]. Furthermore, since the dot-product attention is performed without considering the effect of the influence of features located in different locations. In contrast, early work on image restoration [8] emphasizes the impact of location. Therefore, we propose the Laplace attention, which introduces a new distance prior in the calculation of similarity and represents the effect of spatial location in the form of a Laplace distribution.

In summary, in this paper our contributions are summarized as follows:

- We propose a novel attention-based network (or transformer), called Hourglass Attention Network (HAN) for image inpainting, which combines the respective advantages of attention module and convolution to complete the image features.
- Our proposed hourglass result not only improves the quality of the inpainting image by using hierarchical feature information, but also reduces the computational complexity compared to the vanilla transformer structure.
- We propose Laplace attention, which considers not only the features themselves but also the effect of the distance between features located at different locations when calculating the attention scores. The effect is also more efficient than the position encoding in transformer.
- Experiments on several datasets show that our proposed approach is effective and performs favorably against state of the art inpainting approaches.

2 Related Work

2.1 Image Inpainting

A variety of different approaches have been proposed for image inpainting, and in general these methods can be divided into the following two categories, namely traditional exemplar-based image inpainting methods and learning-based inpainting methods.

Traditional exemplar-based approaches [3,1,7,2,10] usually match and copy background patches into missing areas or by propagating information from boundaries around the missing regions. They perform pretty well on small holes or background inpainting tasks. Nonetheless, due to the low ability to obtain high-level semantic information, they cannot effectively complement images that have complex patterns or generate novel objects that are not present in the observed part.

Learning-based image inpainting approaches usually formulate inpainting as a conditional image generation problem based on CE (Context Encoders) [39], which is the first to introduce a generative adversarial network [14] framework in image inpainting fields and to use an autoencoder as its conditional image generator. Iizuka *et al.* [22] improve the quality of CE by designing a local-global discriminator. Then some researchers [55,46,58,50,62,32] propose a kind of contextual attention module to alleviate the deficiencies of CNNs in capturing long-range dependencies. Next for the problem of spatially sharing parameters, some researchers [30,59,54,53,51] modify convolution operation to adapt the difference between the damaged areas and non-damaged areas in images to obtain more accurate features comparing vanilla convolution. Due to the sparsity of the effective information caused by the broken image, some researchers have tried to guide generation of the missing content through other information of the image, such as edges [37], structure [43,27,31,16]. Finally, to extend the applicability of image inpainting, some researchers have started to focus on high-resolution large-area broken image inpainting [57,63,65], as well as diverse image inpainting [66,64,40,33,49].

2.2 Attention

The attention mechanism can be viewed as a way to bias the allocation of available computational resources towards the most informative components of a signal [20]. The transformer [48] constructed with attention as a cornerstone was firstly proposed for machine translation and has subsequently been proven successful in various down-stream natural language processing tasks. Carion *et al.* [5] started to introduce the transformer to the field of vision, and a series of transformer-based backbone networks [11,29,47,34,52,6,4] for high-level vision tasks were proposed. Moreover, because attention can model dense correlations between input elements well, some models have begun to explore transformer-based models for applications in low-level vision [56,61]. However, existing visual

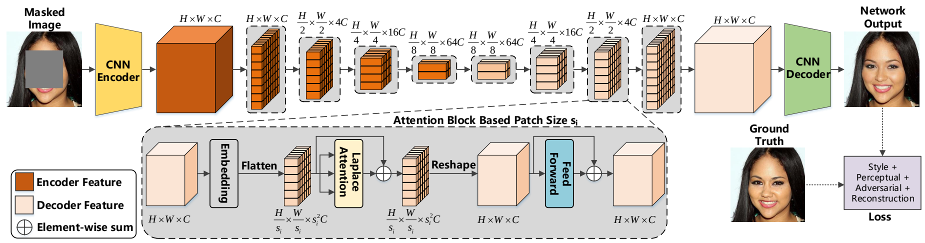

Fig. 2. Pipeline Overview. Our model consists of three parts including a CNN encoder, a CNN decoder, and an hourglass attention structure. The encoder is responsible for extracting features from the input image, and the decoder is used to render the features into an image. The hourglass attention structure is created by stacking designed attention blocks in an hourglass shape, which exploits the powerful long-range modeling capability of attention to fully mine the contextual information in hierarchical features.

transformers often do not focus specifically on the effects of distance between features. Motivated by recent progress in self-attention approaches [15,42,25,18] for language modeling, we propose Laplace attention that remedies this deficiency.

3 Approach

The process of image inpainting is to predict the intact version (ground truth) $\boldsymbol{I}_g$ of a given corrupted image $\boldsymbol{I}_m$ by filling in the missing pixels. The overview of the proposed Hourglass Attention Network (HAN) is shown in Figure 2. HAN contains a CNN encoder, CNN decoder, and the most critical hourglass attention structure. We will describe them in detail below.

3.1 Hourglass Attention Structure

Our hourglass attention structure is composed of tailored attention block based on feature patches of different sizes. The design of this attention block we refer to the encoder block in the vanilla transformer [48] and contains two sublayers. The first is a proposed Laplace attention layer, and the second is a simple feed-forward network (FFN). In addition, we adopt a residual connection [17] adhering to each of the sub-layers. Our hourglass attention structure consists of two processes, followed by feature encoding and feature decoding. In the process of feature encoding, we adopt a gradual reduction strategy to control the number of feature patches, and in the process of decoding, we adopt a gradual increase strategy. Specifically, after the broken image is passed through the encoder, we get a feature map $F \in \mathbb{R}^{H \times W \times C}$. In the process of feature encoding, first we divide F into HW patches, each of size $1 \times 1 \times C$. Then we feed this patch sequence to the sub-layers of the first attention block. The process of the feature passing through the first module is denoted as " Stage E_1".

The procedure is repeated 3 more times with different patch sizes during the process of feature encoding, as "Stage E_2", "Stage E_3" and "Stage E_4". In

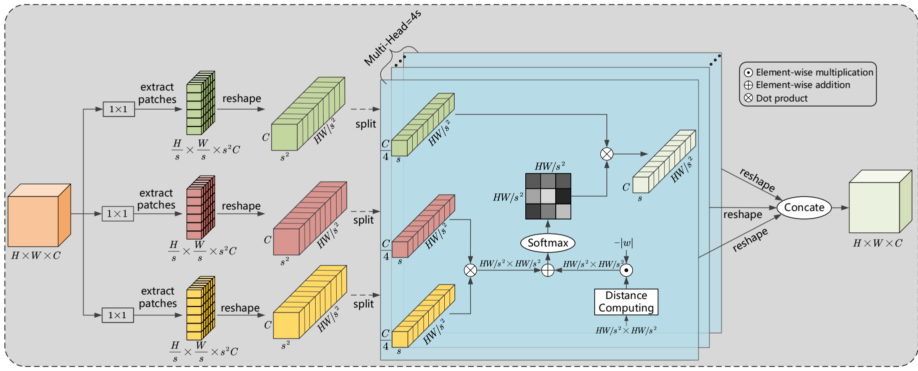

Fig. 3. Details of the proposed Laplace attention. Our proposed laplace prior adds a distance prior to the vanilla multi-head attention to capture the effects of distances located between different spatial locations. "Distance Computing" in the figure means calculating the taxicab geometry between feature patches.

brief in the process of feature encoding, in each "Stage E_i" we firstly divide the input feature map $F \in \mathbb{R}^{H\times W\times C}$ into $\frac{HW}{s_i^2}$ patches, and each of size $s_i \times s_i \times C$ (where $s_i = 1, 2, 4, 8$ in order). Then we feed these feature patches into the subsequent sub-layers of the attention block and output a new feature map with the same size as F. And as the network gets deeper, the number of patches will gradually decrease and the size of patches will gradually increase, which shares some similarities with T2T-ViT [60], PVT [52] and Swin [34]. In addition, the number of patches (spatial resolution) decreases progressively and the dimension of patches (number of channels) increases progressively, which is similar to the classical convolutional network design, such as VGG [45], Resnet [17].

As for the feature decoding process, it can be basically regarded as the inverse process of feature encoding. In each stage of feature decoding, we also divide the input feature $F \in \mathbb{R}^{H\times W\times C}$ into $\frac{HW}{s_i^2}$ patches, with size $s_i \times s_i \times C$ (where $s_i = 8, 4, 2, 1$ in order). Then similar to the feature encoding, these patches are feed to sub-layer of the attention block and obtain the corresponding feature map. The feature decoding process consists of four stages, in the order of "Stage D_4", "Stage D_3", "Stage D_2", and "Stage D_1". And as the network goes deeper, the number of patches (spatial resolution) increases and the dimension of patch (number of channels) decreases similar to classical generative networks, such as DCGAN[41]. These proposed stages are arranged in our inpainting network as "Stage E_1, E_2, E_3, E_4, D_4, D_3, D_2, D_1", which jointly form a symmetric hourglass structure that generates a hierarchical feature representation as a classical autoencoder. They leverage the features information at multiple scales to fully exploit the contextual formation of the input and generate suitable features for the broken ares.

In summary, the hourglass attention structure allows our model to utilize multi-scale information, which not only allows our model to improve performance but also reduces computational complexity, as shown in Table 2 and Table 4.

3.2 Laplace Attention

Our Laplace attention can be regarded as a multi-head self-attention [48] with the special Laplace prior. Suppose that the patch size of the "Stage" where an attention located is s. First we embed a feature map $F \in \mathbb{R}^{H\times W\times C}$ into a query feature $Q \in \mathbb{R}^{H\times W\times C}$, a key feature $K \in \mathbb{R}^{H\times W\times C}$ and a value feature $V \in \mathbb{R}^{H\times W\times C}$ by different linear layers. Then we extract patches of shape $d = s\times s\times C$ from the query Q and we can get $l = H/s\times W/s$ patches. Next we flatten and reshape these patches into column vectors, and then merge the vectors into a matrix $\boldsymbol{Q} \in \mathbb{R}^{d\times l}$, i.e. l d-dimensional patch sequences. Similar operations are performed for key K, value Q to obtain the corresponding $\boldsymbol{K} \in \mathbb{R}^{d\times l}$, $\boldsymbol{V} \in \mathbb{R}^{d\times l}$. Moreover, inspired by the language model [15,42,25,18], we introduce a Laplace prior on the similarity distribution (the softmax output in attention) to reflect the effect of distance in attention. Specifically, suppose the spatial coordinates $\boldsymbol{c}_i$, $\boldsymbol{c}_j$ of patches $\boldsymbol{q}_i$, $\boldsymbol{k}_j$ are (x_i, y_i), (x_j, y_j). For each $\boldsymbol{q}_i$ we introduce a two-dimensional spatial "isotropic" Laplace distribution $p_i(\boldsymbol{c}) \sim \text{Laplace}(\boldsymbol{c} \mid \boldsymbol{\mu}_i, \boldsymbol{I})$ (where $\boldsymbol{\mu}_i = (x_i, y_i)^\top$, and $\boldsymbol{I}$ is an identity matrix) as prior for the attention score (the value obtained after softmax). As shown in the Figure 3, the attention output $\boldsymbol{o}_i \in \mathbb{R}^d$ for i-th query patch $\boldsymbol{q}_i \in \mathbb{R}^d$ in $\boldsymbol{Q}$ can be defined by by:

$$
\begin{aligned}
\boldsymbol{o}_i &= \text{Attention}\left(\boldsymbol{q}_i, \{\boldsymbol{k}_j\}_{j=1}^{l}, \{\boldsymbol{v}_j\}_{j=1}^{l}\right) \\
&= \sum_{j=1}^{l} \frac{p_i(\boldsymbol{c_j}) \exp(\boldsymbol{q}_i^\top \boldsymbol{k}_j)}{\sum_{n=1}^{l} p_i(\boldsymbol{c}_n) \exp(\boldsymbol{q}_i^\top \boldsymbol{k}_n)} \boldsymbol{v}_j \\
&= \sum_{j=1}^{l} \frac{\exp(-t_{ij}) \exp(\boldsymbol{q}_i^\top \boldsymbol{k}_j)}{\sum_{n=1}^{l} \exp(-t_{in}) \exp(\boldsymbol{q}_i^\top \boldsymbol{k}_n)} \boldsymbol{v}_j \\
&= \sum_{j=1}^{l} \text{softmax}_j(q_i^\top k_j - t_{ij}) \boldsymbol{v}_j \\
&\approx \sum_{j=1}^{l} \text{softmax}_j(q_i^\top k_j - |w| t_{ij}) \boldsymbol{v}_j
\end{aligned}
\tag{1}
$$

where $\boldsymbol{k}_j \in \mathbb{R}^d$ is j-th key patch in $\boldsymbol{K}$ and $\boldsymbol{v}_j \in \mathbb{R}^d$ is corresponding j-th value patch in $\boldsymbol{V}$, $1 \leq i, j \leq l$, $t_{ij} = |x_i - x_j| + |y_i - y_j|$ and $|w|$ means a learnable parameter greater than 0. Since the variance of the Laplace distribution $p_i(\boldsymbol{c})$ will not always be $\boldsymbol{I}$ in real situations, we use $|w|$ to represent the variance here to enhance the flexibility of the model. Therefore, the incorporation of the taxicab geometry (l_1-distance) between the patches can also be seen as the Laplace prior when calculating the similarity.

Furthermore, revisiting the process of extracting patches from the feature $Q \in \mathbb{R}^{H\times W\times C}$, when we choose patches of larger size s, the dimension $d = s\times s\times c$ of the patch is also larger, and the length $l = H/s\times W/s$ of the patch sequences is smaller, so the size of the attention matrix $l\times l$ is smaller. Therefore, to alleviate parameter redundancy, we perform $4s$-heads attention in parallel, as:

$$\boldsymbol{O} = \text{Concate}\left(\text{Head }_1, \ldots, \text{Head }_{4s}\right) \tag{2}$$

where $\boldsymbol{O} \in \mathbb{R}^{H\times W\times C}$ and

$$\text{Head }_i = \text{Attention}\left(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i\right)$$

where $\boldsymbol{Q}_i \in \mathbb{R}^{\frac{sC}{4}\times\frac{HW}{s^2}}$, $\boldsymbol{K}_i \in \mathbb{R}^{\frac{sC}{4}\times\frac{HW}{s^2}}$, $\boldsymbol{V}_i \in \mathbb{R}^{\frac{sC}{4}\times\frac{HW}{s^2}}$ are matrix stacked by the vectors $\boldsymbol{q}_i \in \mathbb{R}^{\frac{sC}{4}}$, $\boldsymbol{k}_i \in \mathbb{R}^{\frac{sC}{4}}$, $\boldsymbol{v}_i \in \mathbb{R}^{\frac{sC}{4}}$, respectively.

In summary, with the help of multi-head and the distance prior, the Laplace attention, can effectively borrow relevant features from different regions, which better models the long range dependencies inside feature maps.

3.3 Loss Functions

The loss function L_{all} for training our HAN consists of four terms, containing the L_1 loss, the perceptual loss [23], the style loss [12] and the adversarial loss [14], as:

$$L_{\text{all}} = \alpha L_{\text{re}} + \beta L_{\text{perc}} + \gamma L_{\text{style}} + \lambda L_{\text{adv}} \tag{3}$$

where α, β, γ, and λ hyper-parameters. In our experimental procedure, we set $\alpha = 1$, $\beta = 1$, $\gamma = 250$, and $\lambda = 0.1$.

L_1 Loss The L_1 loss refers to the value of the L_1-norm of the difference between the complementary images $\boldsymbol{I}_{\text{out}}$ and the real image $\boldsymbol{I}_{\text{g}}$, :

$$L_{\text{re}} = \|\boldsymbol{I}_{out} - \boldsymbol{I}_g\|_1 \tag{4}$$

Perceptual Loss The perceptual loss measures the feature map between the real image $\boldsymbol{I}_{\text{g}}$ and the output $\boldsymbol{I}_{\text{out}}$, as:

$$L_{\text{perc}} = \mathbb{E}\left[\sum_i \frac{1}{N_i}\left\|\phi_i\left(\boldsymbol{I}_{out}\right) - \phi_i\left(\boldsymbol{I}_g\right)\right\|_1\right] \tag{5}$$

where ϕ_i is the feature map of the i-th layer of pre-trained VGG-19 [45]. And ϕ_i contains activation Relu1_1 [13], Relu2_1, Relu3_1, Relu4_1, and Relu_51 of the VGG-19.

Style Loss The style loss is similar to perceptual loss, as:

$$L_{\text{style}} = \mathbb{E}_j\left[\left\|G_j^{\Phi}\left(\boldsymbol{I}_{out}\right) - G_j^{\Phi}\left(\boldsymbol{I}_g\right)\right\|_1\right] \tag{6}$$

Where G_j^{Φ} is a $C_j\times C_j$ Gram matrix formed by the corresponding feature maps ϕ_j. Here, ϕ_j contains the same layers as the ϕ_j in perceptual loss.

Table 1. Numerical comparisons on the several datasets. The ↓ indicates lower is better, while ↑ indicates higher is better

DataSet		Paris Street View				Celeba-HQ				Places2			
Mask Ratio		10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%
FID↓	GC	20.68	39.48	58.66	82.51	2.54	4.49	6.54	9.83	18.91	30.97	45.26	61.16
	RFR	20.33	28.93	39.84	49.96	3.17	4.01	4.89	6.11	17.88	22.94	30.68	38.69
	DSN	16.28	29.39	42.02	53.66	1.91	3.18	4.70	6.20	13.64	22.74	31.97	41.14
	DTS	16.66	31.94	47.30	65.44	2.08	3.86	6.06	8.58	15.72	27.88	42.44	57.78
	Ours	**12.39**	**22.70**	**35.29**	**46.93**	**1.49**	**2.58**	**3.93**	**5.39**	**12.01**	**20.15**	**28.85**	**37.63**
PSNR↑	GC	32.28	29.12	26.93	24.80	32.25	29.10	26.71	24.78	28.55	25.22	22.97	21.24
	RFR	30.18	27.76	25.99	24.25	30.93	28.94	27.11	25.47	27.26	24.83	22.75	21.11
	DSN	31.06	28.05	25.92	24.05	32.72	29.53	27.15	25.34	28.39	25.03	22.69	20.97
	DTS	32.69	29.28	26.89	24.97	32.91	29.51	27.02	25.13	28.91	25.36	22.94	21.21
	Ours	**32.97**	**29.92**	**27.60**	**25.67**	**33.04**	**29.94**	**27.53**	**25.62**	**28.93**	**25.44**	**23.06**	**21.38**
SSIM↑	GC	0.960	0.925	0.872	0.800	0.979	0.959	0.931	0.896	0.944	0.891	0.824	0.742
	RFR	0.943	0.908	0.861	0.799	0.970	0.958	0.939	0.913	0.929	0.891	0.830	0.756
	DSN	0.952	0.914	0.859	0.791	0.981	0.963	0.939	0.910	0.946	0.894	0.827	0.749
	DTS	0.963	0.929	0.875	0.812	0.981	0.962	0.937	0.905	0.952	0.901	0.834	0.755
	Ours	**0.966**	**0.936**	**0.891**	**0.834**	**0.983**	**0.967**	**0.945**	**0.918**	**0.957**	**0.903**	**0.839**	**0.762**

Adversarial Loss The adversarial loss is defined by:

$$L_{\text{adv}} = \mathbb{E}_{\boldsymbol{I}_g}\left[\log D\left(\boldsymbol{I}_g\right)\right] + \mathbb{E}_{\boldsymbol{I}_{out}} \log\left[1 - D\left(\boldsymbol{I}_{out}\right)\right] \tag{7}$$

where D is the a PatchGAN discriminator with spectral normalization [36].

4 Experiments

We evaluated our HAN on three public datasets, including Paris street view (Paris) [39], CelebA-HQ [24] and Places2 [67]. For data splitting, in CelebA-HQ we chose the first 2000 images as the test set and the rest as the training set. As for Paris and Places2, we used their original data splitting. The resolution of all images during experiment was resized to 256×256. In addition, we used the classical mask dataset [30] to determine the location of image breakage during the test. Our proposed HAN was implemented based on Pytorch [38]. In the training process we used a RTX3090 (24 GB) and set the batch size to 6. We used an AdamW [35] optimizer with with $\beta_1 = 0.5$, $\beta_2 = 0.9$ to train the model. At the start, a learning rate of 10^{-4} was used to train the model and then we used 10^{-5} for fine-tuning the model. Specifically, on CelebA-HQ and Paris, we trained 600,000 iterations and then fine-tuned 150,000 iterations. On the Places2 data set, we trained about 1.2 million iterations and then fine-tuned 400,000 iterations.

4.1 Baselines

We compare with the following baselines for their state-of-the-art performance:

- GC [59]: a two-stage inpainting model, which leverages the gated convolution and the contextual attention [58].

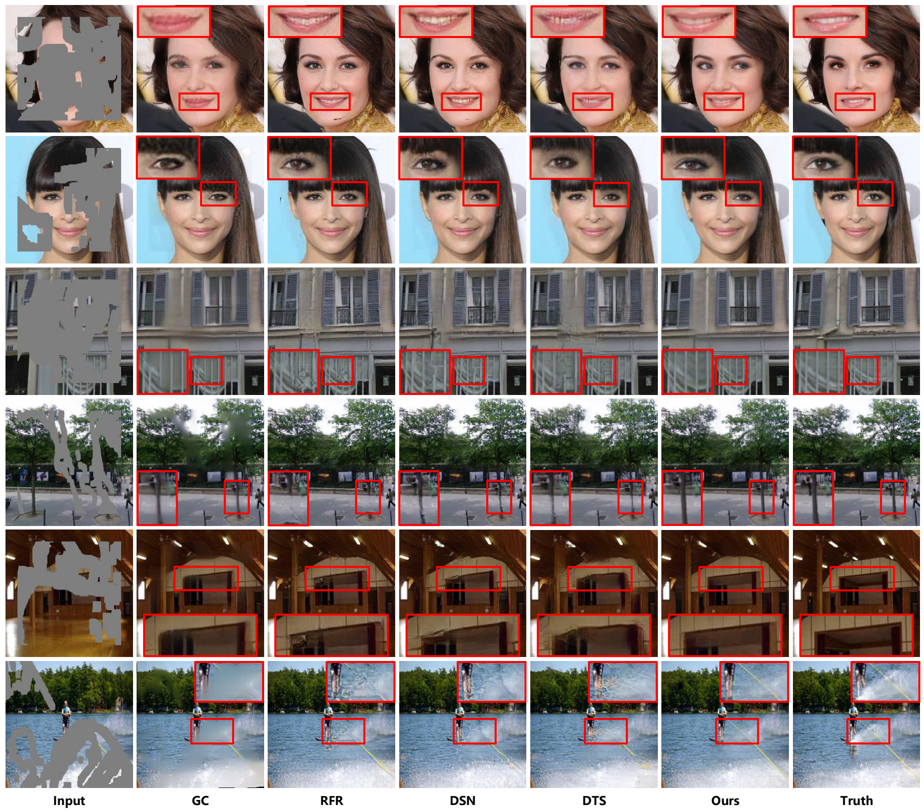

Fig. 4. Qualitative results with GC [59], RFR [28], DSN [51], DTS [16] and our models. The images in each of the two rows from top to the bottom are taken from CelebA-HQ, Paris street view, Places2 respectively. (Best viewed with zoom-in)

- RFR [28]: a recurrent inpainting model with a sepcial contextual attention which recurrently infers the hole and progressively strengthens the result.
- DSN [51]: an U-net inpainting model, which expands the receptive field of convolution based on deformable convolution [9] to skip those broken features and thus learns more valid information.
- DTS [16]: a dual U-net inpainting model, which recovers corrupted images by simultaneous modeling structure-constrained texture synthesis and texture-guided structure reconstruction.

4.2 Quantitative Comparison

We chose FID (Fréchet Inception Distance) [19], PSNR (peak signal-to-noise ratio), and SSIM (structural similarity index) to evaluate our model. SSIM and PSNR measure the similarity of pixels and structural information from paired

Fig. 5. Fail results from Places2 with GC [59], RFR [28], DSN [51], DTS [16] and our models. when a large portion of the image got corrupted, our method is not able to obtain sufficient long term dependency information to assist reconstruction, and hence only the small handset wire part got restored. (Best viewed with zoom-in)

images. SSIM and PSNR are widely used for image evaluation and measure the similarity of pixels and structural information from paired images to provide an appropriate approximation to human visual perception. Nonetheless, sometimes the results of inpainting are diverse from original images for the target areas (e.g. object removal described in [59]), while these metrics are limited to comparing with the original image content. Therefore we also adopted FID to indicate the perceptual quality of the results as generally adopted metric in image generation. As seen from Table 1, our proposed model achieves superior results compared with other baselines in almost all metrics. Meanwhile, favorable performance is achieved in our proposed method of filling irregular holes with various hole versus image ratios. It is worth noting that the advantage of our model tends to be more pronounced when the percentage of breakage is larger compared to other methods, which demonstrates the stronger adaptability of our proposed method to inputs mixed with invalid information with the addition of the hourglass attention structure.

4.3 Qualitative Comparisons

Figure 4 shows qualitative results with previous state-of-art baselines to ours. GC [59] can get pretty credible results, but there is still some blurring on the completed images. The images predicted by RFR [28] are quite good in terms of detail texture, but the downside is that some artifacts appear on the generated images. The results generated by DSN [51] show fewer artifacts compared to RFR but are still not particularly desirable. The image produced by DTS [16] basically has no obvious artifacts, but when it recovers an image with complex patterns, the content filled is often not consistent with the original image, e.g., double eyelid on the left but single eyelid on the right in the second row of the Figure 4. In contrast, our method generally does not bring significant artifacts when completing the image, and learns to represent structures and textures in a consistent formation. Note the last row of Figure 4, when part of the stick can be observed, our reconstructed portion shows a well connected stick structure, which also show the effectiveness by our method when non-local information can be modeled. In addition we show cases of failure of each model, as shown in Figure 5. More qualitative comparisons are shown in the supplementary material.

Table 2. Ablation study on impact of the hourglass structure. The ↓ indicates lower is better, while ↑ indicates higher is better

	FID↓				PSNR↑				SSIM↑			
Mask Ratio	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%
w/ reverse	12.93	23.84	36.63	48.56	32.45	29.65	27.43	25.39	0.963	0.933	0.886	0.828
w/o hourglass	13.13	24.42	36.18	49.33	32.53	29.59	27.38	25.37	0.963	0.933	0.887	0.829
Ours	**12.39**	**22.70**	**35.29**	**46.93**	**32.97**	**29.92**	**27.60**	**25.67**	**0.966**	**0.936**	**0.891**	**0.834**

Table 3. Ablation study about the layer number of the hourglass attention. The ↓ indicates lower is better, while ↑ indicates higher is better

	FID↓				PSNR↑				SSIM↑			
Mask Ratio	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%
+0	16.08	32.22	48.27	68.21	31.66	28.63	26.07	23.87	0.956	0.918	0.854	0.771
+1	13.33	24.75	37.84	51.31	32.62	29.59	27.20	25.17	0.963	0.931	0.882	0.819
+2	13.09	23.69	36.55	49.71	32.73	29.71	27.43	25.26	0.964	0.933	0.886	0.824
+3	12.99	23.12	35.51	47.18	32.75	29.82	27.55	25.37	0.964	0.934	0.890	0.827
+4(ours)	**12.39**	**22.70**	**35.29**	**46.93**	**32.97**	**29.92**	**27.60**	**25.67**	**0.966**	**0.936**	**0.891**	**0.834**

5 Ablation Study

We explore the impact of our proposed module on the Paris dataset.

5.1 Effectiveness of Hourglass Attention Structure

Here we validate the role of the proposed hourglass attention structure, and the results are shown in Tables 2 and 3 respectively. In Table 2, we design two other attention structures to compare with our hourglass structure, including the standard structure (similar to the vanilla transformer, with patch size of the attention module all set to s=1, i.e. without hourglass structure) and the spindle structure (i.e., reversing the order of the attention blocks in hourglass structure, as "Stage E_4, E_3, E_2, E_1, D_1, D_2, D_3, D_4", denoted by "reverse"). As we can see, compared to the standard structure, our hourglass structure utilizes multi-scale hierarchical feature information more helpful for image inpainting. Additionally, the hourglass structure from smallest to largest during encoding and largest to smallest during decoding is also more reasonable than the spindle structure with reversed order.

Further, we performed a series of experiments to demonstrate the effectiveness of hierarchical attention, as shown in Table 3. In the Table 3, S_0 represents no inclusion attention module, S_1 represents inclusion only "Stage E_1, D_1", then S_2 represents inclusion "Stage E_1, E_2, D_2, D_1", and so on. We find that stacking more hierarchical attention can bring continuous improvements.

Finally, we show the advantage of the hourglass structure in terms of complexity and compare it with other baseline models, as shown in Table 4, where "MHA" represents the multi-headed attention (i.e., the vanilla transformer) without the hourglass structure. For a feature $F \in \mathbb{R}^{H\times W\times C}$, the complexity of attention can be simplified to $\mathcal{O}(CH^2W^2/s^2)$, where s is the size of the

Table 4. Model complexity. Here we provide the FLoating-point OPerations (FLOPs) and parameters (Params) of the model

Model	GC	RFR	DSN	DTS	MHA	Ours
FlOPs	103.1G	206.1G	24.8G	75.9G	183.6G	137.7G
Params	16.0M	30.6M	99.3M	52.1M	19.4M	19.4M

Table 5. Ablation study on impact of the distance prior. The ↓ indicates lower is better, while ↑ indicates higher is better

	FID↓				PSNR↑				SSIM↑			
Mask Ratio	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%
w/ Gus	12.94	23.50	35.38	48.26	32.90	29.81	27.52	25.59	0.965	0.934	0.889	0.831
w/ Sin	13.06	23.55	36.39	48.17	32.79	29.76	27.52	25.52	0.964	0.934	0.889	0.832
w/o Lap	13.45	24.57	36.49	48.67	32.50	29.56	27.34	25.42	0.962	0.931	0.885	0.827
Ours	**12.39**	**22.70**	**35.29**	**46.93**	**32.97**	**29.92**	**27.60**	**25.67**	**0.966**	**0.936**	**0.891**	**0.834**

patch. It can be seen that due to the presence of the hourglass dividing the patches of larger size ($s = 1$ in the vanilla transformer), our hourglass structure improves the performance and reduces the complexity at the same time. In addition, it can be seen from the table that our inpainting model is able to maintain fewer model parameters and moderate computational effort while achieving a performance lead compared to other baseline models.

5.2 Effectiveness of distance prior

In our model we introduce a two-dimensional Laplace prior to represent the effect of the distance between feature patches located in different regions. On the other hand, many visual transformers [5,52] tend to represent this influence using an extension of the one-dimensional position encoding (of the vanilla transformer [48]) to two dimensions. Here we replace our proposed distance prior with a 2D position encoding (implemented by trigonometric functions, denoted "Tri") and the results show that the position encoding is less effective than the proposed distance prior, as shown in Table 5. Furthermore, we also compare the case of replacing the Laplace prior with a Gaussian prior (i.e. replacing the l_1-distance with a square of l_2-distance, denoted by "Gus"), removing the distance prior (denoted by "no"), and removing the variance coefficient $|w|$. From the table 5, it can be seen that the Laplace prior, which makes the attention score decrease more slowly, is more suitable for image completion than the Gaussian prior. This may show that for inpainting, with features farther away from the target can still provide certain valid information for the target. Besides the variance coefficient $|w|$ makes the model more robust.

6 Conclusion

In this paper, we propose to learn the hourglass attention network (HAN) for image mapping, which builds an hourglass attention structure based on the powerful texture pattern learning capability of CNNs to mine the contextual information in hierarchical features to synthesize appropriate contents for the complemented images. Besides, we introduce a new distance prior to the attention mechanism, making the attention to consider not only the similarity of the features themselves, but also the influence of distance between the features. Quantitative and qualitative results show that our model is capable of generating more coherent and fine-detailed results.

Limitation Similar to other learning-based inpainting models [59,28,51,16], it is still difficult for our HAN to handle images that have complex patterns suffering from extreme large breakage ratios.

Broader Impact The proposed method will reflect the biases of the datasets they are trained on and may generate inexistent content. If deployed without careful consideration, inpainting methods (including but not limited to HAN) trained on research datasets like Celeba-HQ and Places2 may bring negative affect by propagating biases in the dataset. These issues warrant further research and consideration when building upon this work.

Acknowledgments

This work is jointly supported by the National Key Research and Development Program of China under Grant No. 2017YFA0700800, the General Program of China Postdoctoral Science Foundation under Grant No. 2020M683490, and the Youth program of Shaanxi Natural Science Foundation under Grant No. 2021JQ-054.

References

1. Ballester, C., Bertalmio, M., Caselles, V., Sapiro, G., Verdera, J.: Filling-in by joint interpolation of vector fields and gray levels. IEEE Transactions on Image Processing **10**(8), 1200–1211 (2001). https://doi.org/10.1109/83.935036
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: A randomized correspondence algorithm for structural image editing. ACM Transactions on Graphics (Proc. SIGGRAPH) **28**(3) (Aug 2009)
3. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques. p. 417–424. SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., USA (2000). https://doi.org/10.1145/344779.344972
4. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. arXiv preprint arXiv:2105.05537 (2021)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 213–229. Springer International Publishing, Cham (2020)
6. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
7. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Transactions on image processing **13**(9), 1200–1212 (2004)
8. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-d transform-domain collaborative filtering. IEEE Transactions on Image Processing **16**(8), 2080–2095 (2007). https://doi.org/10.1109/TIP.2007.901238
9. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
10. Ding, D., Ram, S., Rodríguez, J.J.: Image inpainting using nonlocal texture matching and nonlinear filtering. IEEE Transactions on Image Processing **28**(4), 1705–1719 (2019). https://doi.org/10.1109/TIP.2018.2880681
11. 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: International Conference on Learning Representations (2021)
12. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
13. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Gordon, G., Dunson, D., Dudík, M. (eds.) Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 15, pp. 315–323. PMLR, Fort Lauderdale, FL, USA (11–13 Apr 2011)
14. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014)

15. Guo, M., Zhang, Y., Liu, T.: Gaussian transformer: A lightweight approach for natural language inference. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 6489–6496 (Jul 2019)
16. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14134–14143 (October 2021)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
18. He, P., Liu, X., Gao, J., Chen, W.: {DEBERTA}: {DECODING}-{enhanced} {bert} {with} {disentangled} {attention}. In: International Conference on Learning Representations (2021)
19. 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: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
20. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
21. Huang, J.B., Kang, S.B., Ahuja, N., Kopf, J.: Image completion using planar structure guidance. ACM Transactions on graphics (TOG) **33**(4), 1–10 (2014)
22. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and Locally Consistent Image Completion. ACM Transactions on Graphics (Proc. of SIGGRAPH 2017) **36**(4), 107:1–107:14 (2017)
23. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), `https://openreview.net/forum?id=Hk99zCeAb`
25. Ke, G., He, D., Liu, T.Y.: Rethinking positional encoding in language pre-training. In: International Conference on Learning Representations (2021)
26. Komodakis, N., Tziritas, G.: Image completion using efficient belief propagation via priority scheduling and dynamic pruning. IEEE Transactions on Image Processing **16**(11), 2649–2661 (2007). https://doi.org/10.1109/TIP.2007.906269
27. Li, J., He, F., Zhang, L., Du, B., Tao, D.: Progressive reconstruction of visual structure for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
28. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
29. Li, Y., Zhang, K., Cao, J., Timofte, R., Van Gool, L.: Localvit: Bringing locality to vision transformers. arXiv preprint arXiv:2104.05707 (2021)
30. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
31. Liu, H., Jiang, B., Song, Y., Huang, W., Yang, C.: Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In: Computer Vision – ECCV 2020. pp. 725–741. Springer International Publishing, Cham (2020)

32. Liu, H., Jiang, B., Xiao, Y., Yang, C.: Coherent semantic attention for image inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
33. Liu, H., Wan, Z., Huang, W., Song, Y., Han, X., Liao, J.: Pd-gan: Probabilistic diverse gan for image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9371–9381 (June 2021)
34. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10012–10022 (October 2021)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), `https://openreview.net/forum?id=Bkg6RiCqY7`
36. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018), `https://openreview.net/forum?id=B1QRgziT-`
37. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: Edgeconnect: Structure guided image inpainting using edge prediction. In: The IEEE International Conference on Computer Vision (ICCV) Workshops (Oct 2019)
38. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
39. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
40. Peng, J., Liu, D., Xu, S., Li, H.: Generating diverse structure for image inpainting with hierarchical vq-vae. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10775–10784 (June 2021)
41. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
42. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research **21**(140), 1–67 (2020)
43. Ren, Y., Yu, X., Zhang, R., Li, T.H., Liu, S., Li, G.: Structureflow: Image inpainting via structure-aware appearance flow. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
44. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
45. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)

46. Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Kuo, C.C.J.: Contextual-based image inpainting: Infer, match, and translate. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
47. Vaswani, A., Ramachandran, P., Srinivas, A., Parmar, N., Hechtman, B., Shlens, J.: Scaling local self-attention for parameter efficient visual backbones. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12894–12904 (June 2021)
48. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
49. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4692–4701 (October 2021)
50. Wang, N., Li, J., Zhang, L., Du, B.: Musical: Multi-scale image contextual attention learning for inpainting. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. pp. 3748–3754. International Joint Conferences on Artificial Intelligence Organization (7 2019). https://doi.org/10.24963/ijcai.2019/520
51. Wang, N., Zhang, Y., Zhang, L.: Dynamic selection network for image inpainting. IEEE Transactions on Image Processing **30**, 1784–1798 (2021). https://doi.org/10.1109/TIP.2020.3048629
52. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 568–578 (October 2021)
53. Wang, Y., Chen, Y.C., Tao, X., Jia, J.: Vcnet: A robust approach to blind image inpainting. In: Computer Vision – ECCV 2020. pp. 752–768. Springer International Publishing, Cham (2020)
54. Xie, C., Liu, S., Li, C., Cheng, M.M., Zuo, W., Liu, X., Wen, S., Ding, E.: Image inpainting with learnable bidirectional attention maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
55. Yan, Z., Li, X., Li, M., Zuo, W., Shan, S.: Shift-net: Image inpainting via deep feature rearrangement. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
56. Yang, F., Yang, H., Fu, J., Lu, H., Guo, B.: Learning texture transformer network for image super-resolution. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
57. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
58. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
59. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
60. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z.H., Tay, F.E., Feng, J., Yan, S.: Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 558–567 (October 2021)

61. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 528–543. Springer International Publishing, Cham (2020)
62. Zeng, Y., Fu, J., Chao, H., Guo, B.: Learning pyramid-context encoder network for high-quality image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
63. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 1–17. Springer International Publishing, Cham (2020)
64. Zhao, L., Mo, Q., Lin, S., Wang, Z., Zuo, Z., Chen, H., Xing, W., Lu, D.: Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
65. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I.C., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. In: International Conference on Learning Representations (2021)
66. Zheng, C., Cham, T.J., Cai, J.: Pluralistic image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
67. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1452–1464 (2018). https://doi.org/10.1109/TPAMI.2017.2723009