Multimedia Systems (2024) 30:21
https://doi.org/10.1007/s00530-023-01211-w

REGULAR PAPER

ITrans: generative image inpainting with transformers

Wei Miao[1,4] · Lijun Wang[2] · Huchuan Lu[1] · Kaining Huang[3] · Xinchu Shi[3] · Bocong Liu[3]

Received: 15 June 2023 / Accepted: 8 December 2023 / Published online: 17 January 2024
© The Author(s) 2024

Abstract
Despite significant improvements, convolutional neural network (CNN) based methods are struggling with handling long-range global image dependencies due to their limited receptive fields, leading to an unsatisfactory inpainting performance under complicated scenarios. To address this issue, we propose the Inpainting Transformer (ITrans) network, which combines the power of both self-attention and convolution operations. The ITrans network augments convolutional encoder–decoder structure with two novel designs, i.e. , the global and local transformers. The global transformer aggregates high-level image context from the encoder in a global perspective, and propagates the encoded global representation to the decoder in a multi-scale manner. Meanwhile, the local transformer is intended to extract low-level image details inside the local neighborhood at a reduced computational overhead. By incorporating the above two transformers, ITrans is capable of both global relationship modeling and local details encoding, which is essential for hallucinating perceptually realistic images. Extensive experiments demonstrate that the proposed ITrans network outperforms favorably against state-of-the-art inpainting methods both quantitatively and qualitatively.

Keywords Convolutional neural network · Image inpainting · Global transformer · Local transformer

1 Introduction

Image inpainting, also known as image completion, is the task of filling in missing pixels in an image with fine image content. This task finds applications in various image editing domains, including object removal [1], image restoration [2], photo retouching [3], etc. The solution of image inpainting is to understand image structures and perform image synthesis. Prior to the deep learning era, this subject was mainly performed by using existing image patches to fill in masked regions [2, 4, 5]. However, due to the lack of semantic understanding, these methods have been replaced by deep neural networks [6–14] and adversarial learning [15–19]. Deep learning-based methods treat inpainting as a generation task that involves end-to-end learning using convolutional neural networks (CNNs). CNNs are known for their remarkable capacity to generate fine details, thanks to their inductive biases [20], including locality and weight sharing, which ensure them efficient models across domains. Nevertheless, the limited receptive fields of CNNs are insufficient to access the necessary information for generating quality inpaintings

Communicated by B. Bao.

✉ Wei Miao
mdd98@mail.dlut.edu.cn

Lijun Wang
ljwang@dlut.edu.cn

Huchuan Lu
lhchuan@dlut.edu.cn

Kaining Huang
huangkaining@meituan.com

Xinchu Shi
shixinchu@meituan.com

Bocong Liu
liubocong@meituan.com

1 School of Information and Communication Engineering, Dalian University of Technology, No. 2 Linggong Road, Dalian 116023, Liaoning, China

2 School of Artificial Intelligence, Dalian University of Technology, No. 2 Linggong Road, Dalian 116023, Liaoning, China

3 Meituan Group, No. 4 Wangjing East Road, Chaoyang District, Beijing 100102, China

4 Faculty of Information Technology, University of Jyväskylä, Seminaarinkatu 15, 40014 Jyväskylä, Finland

under complex scenarios, leading to unwanted artifacts and blurry results. More recently, transformers [21] have demonstrated record-breaking performance in various computer vision tasks with strong capability in modeling long-range dependencies. While transformers provide an alternative to CNNs, their lack of inductive biases presents a challenge for processing images. Although transformers have a higher performance ceiling than CNNs, their complex costly pre-training requirements make them more difficult to learn [20]. As a result, using transformers for image inpainting is still relatively uncommon in literature.

We propose the Inpainting Transformer (ITrans) network to integrate the benefits of both CNNs and transformers. To leverage the inductive bias of CNNs, we design a convolutional encoder–decoder network for feature extraction and image generation, respectively. Additionally, we introduce a global and local transformer module to enhance the flexibility of transformers. The global transformer module aims to achieve high-level perception of the input image from a holistic view and connects the encoder with decoder via skip paths in a multi-scale manner. The local transformer module is designed to ensure image consistency and enhance local details at a lower computational cost. Our ITrans network incorporates these novel designs, resulting in greater representational power than pure CNNs and more efficient learning than pure transformers. Consequently, our approach achieves superior performance on various image inpainting benchmarks. Our main contributions are outlined below:

- We propose ITrans network for image inpainting, which benefits from the built-in inductive biases of CNNs and the strong expressive power transformers. And ITrans is the first approach to train CNN-transformer in an end-to-end way.
- We design the global and local transformer module, which learn to capture image context from multiple perspectives and significantly improve image generation of missing regions.
- Our method sets new state-of-the-art performance on various benchmark datasets. Extensive experiments also verify the major insight of our ITrans network.

To the best of our knowledge, we are among the first to investigate hybrid architectures for image inpainting by merging CNNs and transformers in an end-to-end scheme. Our source code and models will be made available upon acceptance.

2 Related work

2.1 Image inpainting

Deep generative networks for image inpainting Traditional patch-based inpainting methods typically rely on propagating images from remaining areas or other sources. For instance, in [2], redundant image patches were employed to determine the priority of each pixel based on gradient variation. Pixels sharing greater similarity with missing pixels were used to fill in the areas. Hays et al. [4] searched through numerous image patches on the Internet to locate a suitable patch to fill missing areas. Another typical patch-based technique is patch matching [5], which looked for identical patches in various source pictures. This method split the image into small patches and selected the most comparable one to fill in the holes. Although these traditional methods work well for small, tiny holes, and homogeneous background regions, they lack the crucial generating capacity needed to handle massive missing regions.

Pathak et al. [6] proposed a deep-learning approach called context encoders for image inpainting tasks, which is the pilot study of this area. Built on an encoder–decoder architecture, the encoder extracts low-resolution features from the corrupted image, and the decoder enlarges and reconstructs the image. However, the approach often results in visual artifacts and blurriness in the recovery regions. To address this issue, Iizuka et al. [7] reduced the number of downsampling layers, and [22] included dilated convolution layers in the bottleneck. Meanwhile, recent work LaMa [23] employed Fourier convolutions to enlarge the receptive filed and inductive bias. The U-Net structure [24] is widely applied to extract low-level features well-reserved in encoder layers. Liu et al. [8] introduced partial convolution to prevent the feature maps from capturing too many zeros, thereby smoothing the output image by filtering redundant zeros while traversing over missing regions. Additionally, Yu et al. [11] implemented gated convolution in both encoder and decoder layers, which learns a dynamic feature selection mechanism for channel-wise spatial placement across all layers, improving color consistency and inpainting quality on free-form masks.

Attention mechanism Attention mechanisms have recently been applied to improve inpainting tasks. Yu et al. [25] first introduced contextual attention, demonstrating the attention process with dilated convolutions. The contextual attention model operates in two stages. The first stage generates a coarse inpainting result, while the second stage refines the image using patch-similarity-based contextual attention. Kim et al. [26] made further progress by introducing the texture transform attention (TTA) module.

With the TTA module, high-level features are reassembled from low-level features and sent to the decoder, improving the inclusion of texture information in the reconstructed regions. Transformers are all about attention, and the inclusion of an attention mechanism reminds us of the potential of transformers in image inpainting.

2.2 Vision transformers

Transformers Attention-based models, particularly transformers [21] have emerged as the de facto standard approach in natural language processing (NLP) [27]. Transformers also show great performance in computer vision fields. Vision transformer (ViT) [28] is a convolution-free transformer that outperforms previous CNN-based models [29] in image recognition tasks. ViT processes images as a sequence of 16×16 words, allowing for robust representation. The transformer architecture's effectiveness in ViT has been demonstrated through pre-training on massive datasets. Subsequently, DeiT [30] adapts ViT for better sample efficiency through an innovative knowledge distillation technique. ViT has also been applied to other computer vision tasks, such as object detection [31, 32], and semantic segmentation [33–35].

Transformers in inpainting Image inpainting can be considered as a form of image generation task, and there are various approaches that use transformers and convolution layers. Parmer et al. [36] first suggested that image generation be viewed as autoregressive sequence generation using a transformer architecture. Generative models such as [36, 37] employ autoregressive learning and GPT-3-based techniques [27]. In contrast, transformer-based generative adversarial networks (GANs) have not received much attention until recently. TransGAN [38] introduced a pure transformer-based GAN that employs grid self-attention, a variant of self-attention, to scale for varying image sizes. ViTGAN [39] modified the normalization layers and output mapping layers of ViT in the encoder to fill in missing regions. To ensure Lipschitzness, ViTGAN utilizes L2 attention [40].

In the field of image inpainting, [41] employed a GPT-based [42] bilateral transformer as the bottleneck model, with convolution-based encoder and decoder for feature extraction. The bilateral transformer is applied to non-predicted tokens, while an autoregressive model is used in predicted tokens to avoid information leakage. This enables the model to simultaneously obtain bilateral context and generate output. ICT [43] is a pluralistic image completion model that consists of two stages. In the first stage, a bi-directional transformer is used to generate a probability distribution for the missing regions. In the second stage, a guided upsampling network is employed to reconstruct the images. T-Fill [44] employed a restrictive CNN for individual weighted token representation, which is used in long-range transformer interactions. Notably, to the best of our knowledge, no other works have attempted to apply the ViT structure in image inpainting.

3 Approach

Our goal is to generate a realistic image $\mathbf{I}_p$ from a masked image $\mathbf{I}_m$ that has missing regions, indicated by a binary mask $\mathbf{M}$. Following the idea of [9], we divide the inpainting process into two stages: edge generation and image inpainting. Specifically, we first generate an edge map with the Canny edge detector [45] and complete the edge map as the image's structure prior. Subsequently, we stack edge map and masked image $\mathbf{I}_m$ together as a four-channel input to our ITrans network to yield the inpainting result.

Our ITrans network is an end-to-end network that incorporates the CNN network with transformer modules, global and local transformer, to enhance the quality of inpainting. Figure 2 shows the structure of the ITrans network. The global transformer focuses on global high-level context modeling in the encoder, and serves in the skip layer to enhance inpainting performance. The local transformer is applied to an additional neighborhood branch to acquire low-level details. These two transformer modules introduce additional attention in inpainting. The network is trained on places and human face datasets with randomly generated irregular masks for free-form inpainting. The architecture details of our ITrans network will be discussed in depth in Sect. 3.1. The two transformer modules in ITrans will be introduced in Sects. 3.2 and 3.3, respectively. The loss functions will be shown in Sect. 3.4.

3.1 ITrans network

The whole inpainting network consists of two stages: edge completion and image inpainting. We have utilized the same edge generation model as [9]. The edge completion model takes the masked grayscale image $\mathbf{I}_g$, masked edge $\mathbf{E}_m$, and Canny-generated edges together as the input to construct the full edge, considered as an image structure prior.

In the image inpainting stage, we introduce ITrans network. ITrans's primary structure is a CNN-based encoder–decoder network, with 8 ResBlocks [29] used to generate missing pixels in the bottleneck. The architecture of ITrans leverages the inductive bias of CNN networks

to efficiently learn cross-domain information from various images. Since encoder features typically contain more unique image structures than decoder features, we believe it is essential to aggregate both types of features for inpainting. To achieve this, we employ the global transformer in the skip-connection structure to merge these two features.

Moreover, we also incorporate an extra branch with four convolutional layers for the local transformer. This feature map is then passed through the local transformer to extract local details. Finally, the concatenation of the ResBlock bottleneck and local transformer outputs is sent to the decoder. This decoder progressively upsamples the feature maps to generate the final image.

The ITrans network is a generative model that trains under the GAN framework [46], using PatchGAN [47] structure for the discriminators. In particular, we choose differnent normalization approaches in different modules. Spectral normalization [48] is applied in all discriminators to stabilize training by scaling down weight matrices. Instance normalization [49] is used in the encoder and decoder for structure generating, while layer normalization [50] is implemented in all transformer layers.

3.2 Global transformer

The global transform00er performs global self-attention on feature maps in order to enhance the quality of image inpainting. Based on the concept of vision transformer (ViT) [51], which treats images as word sequences in natural language processing, our global transformer splits the input image into fixed-size patches and adds class tokens into patches. Position embedding is then used to maintain positional information, and the concatenated sequence is sent to the transformer encoder. To avoid overfitting, dropout layers [52] are also implemented.

In image inpainting tasks, retaining all pixels throughout the process is necessary to preserve key textural clues in the background. Therefore, in the global transformer for image inpainting, we remove all dropout layers to maintain all features and pixels for higher inpainting quality, applied in both position embedding and the transformer encoder. To induce self-attention of the input, a multi-head layer (MLP) is inserted after the transformer encoder. The MLP layer enhances the generation performance of the global transformer and stabilizes training. Following that, a classification vector is obtained, representing categories of all pixels in feature maps. However, instead of a 1-D vector, we want a self-attention map for the decoder. The obtained vector is sent to a rearrange module, which reshapes it into the size of input feature map. Each pixel in this self-attention map has a classification. Finally, we add a convolution layer to recover input channels and smoothing the attention map. And this is the output of the global transformer, which comprises classification categories from the input feature map. The structure of our global transformer is depicted in Fig. 3.

3.3 Local transformer

In general, convolution layers focus on the local area within the convolution kernel, while the ViT module concentrates on global attention and precise details on local areas. However, the global receptive field of ViT can result in the loss of some details. Therefore, to address this issue, we propose the local transformer, which primarily concerns low-level image details in deeper layers. To the best of our knowledge, there have been relatively few attempts to use transformers to extract local fine details. The structure of our local transformer is depicted in Fig. 4.

Initially, we consider the sequences for attention computation, which are query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$). We apply a sequence extracting convolution layer instead of patching procedures to obtain the sequences. The sequences are defined as:

$$\begin{aligned} \mathbf{Q} &= Reshape(f(\mathbf{X})), \\ \mathbf{K} &= Reshape(g(\mathbf{X})), \\ \mathbf{V} &= Reshape(h(\mathbf{X})), \end{aligned} \tag{1}$$

where $\mathbf{X}$ is the input feature map; $f(\cdot)$, $g(\cdot)$, and $h(\cdot)$ are different convolution layers. Then, query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$) sequences are sent to the kernel-sized self-attention layer. The self-attention layer in our local transformer focuses on attention with a sequence size that extracts convolution kernels. To ensure efficient computation, we adopt a dynamic multi-headed dimension choosing mechanism in the attention layer. For multi-headed layers, we use different head numbers for distinct feature channels to save computational costs. The number of head dimensions depends on the number of input feature channels. The head dimension is small for low-level features and large for high-level features, resulting in reduced computation costs across a spectrum of input sizes. The self-attention head is defined as:

$$Attention(\mathbf{X}) = softmax(\mathbf{Q}\mathbf{K}^T/\sqrt{d_h})\mathbf{V}, \tag{2}$$

where $\sqrt{d_h}$ is the feature dimension for each head. Finally, an MLP layer is added to restore the missing pixels and to generate the final local attention map.

To reduce computational costs, we omit the use of position embedding and class tokens in our local transformer design. Attention sequences are generated using convolution kernels, which preserve the order of original features. Therefore, it is unnecessary to retain position information through position embedding. In highly detailed contexts, there are more pixel categories than in the original image, and class

tokens become less significant in local areas while consuming more time. To address this issue, we add a skip layer to the local transformer, which combines the input feature map with the local attention map for upsampled decoding. The output of our local transformer is defined as:

$$Output = F(Concat(\mathbf{X}, Attention(\mathbf{X}))), \tag{3}$$

where $F(\cdot)$ denotes convolution operation.

3.4 Training losses

Inpainting tasks are inherently ambiguous, especially when dealing with extensive missing regions, and multiple plausible fillings may be appropriate for the same region. To address the complexity of this task, we will introduce all of our proposed losses.

In the edge completion stage, we apply adversarial loss and feature-matching loss [53]:

$$\min_{G_1}\max_{D_1} L_{G_1} = \min_{G_1}(\lambda_{adve}\max_{D_1}(L_{adve}) + \lambda_{FM}L_{FM}). \tag{4}$$

The loss weight λ_{adve} and λ_{FM} are to 1 and 10, respectively. Adversarial loss ensures the generated details are naturally looking ones, which is defined as:

$$\begin{aligned} L_{adve} = &\ \mathbb{E}_{(\mathbf{E}_{GT},\mathbf{I}_g)} \log D_1(\mathbf{E}_{GT}, \mathbf{I}_g) \\ &+ \mathbb{E}_{\mathbf{I}_g} \log [1 - D_1(\mathbf{E}_p, \mathbf{I}_g)], \end{aligned} \tag{5}$$

where G_1 and D_1 denote edge generator and discriminator, respectively; $\mathbf{E}_{GT}$ indicates ground truth edges; $\mathbf{E}_p$ indicates predicted completed edges; and $\mathbf{I}_g$ indicates grayscale images.

Feature-matching loss compares activation maps in specific discriminator layers, which is similar to perception loss [54–56] and it is defined as:

$$L_{FM} = \mathbb{E}\left[\sum_{i=1}^{L} \frac{1}{N_i} \|D_1^{(i)}(\mathbf{E}_{GT}) - D_1^{(i)}(\mathbf{E}_p)\|_1\right], \tag{6}$$

where L is the final convolution layer of discriminator, N_i is the number of elements of the i'th activation layer, and $D_1^{(i)}$ is the i'th layer of discriminator.

In inpainting stage, the input is incomplete image $\mathbf{I}_m = \mathbf{I}_{GT} \cdot (1 - \mathbf{M})$, where masked areas are set to 0, along with completed edge map $\mathbf{E}_c = \mathbf{E}_{GT} \cdot (1 - \mathbf{M}) + \mathbf{E}_p \cdot \mathbf{M}$. The predicated image $\mathbf{I}_p$ is generated from the incomplete image and the completed edge. L1 loss, adversarial loss, style loss, perceptual loss, and total variation loss are all included in training loss. L1 loss is normalized by mask size to guarantee a proper scaling. Adversarial loss is similar to Eq. (5):

$$\begin{aligned} L_{advi} = &\ \mathbb{E}_{(\mathbf{I}_{GT},\mathbf{E}_c)} \log D_2(\mathbf{I}_{GT}, \mathbf{E}_c) \\ &+ \mathbb{E}_{\mathbf{E}_c} \log [1 - D_2(\mathbf{I}_p, \mathbf{E}_c)]. \end{aligned} \tag{7}$$

Perceptual loss [54] evaluates the distance between features of the predicted and original images on a pre-trained network. It does not require the exact reconstruction, allowing for variances in the predicted image. Perceptual loss is defined as:

$$L_p = \mathbb{E}\left[\sum_i \frac{1}{N_i} \|\Phi_i(\mathbf{I}_{GT}) - \Phi_i(\mathbf{I}_p)\|_1\right], \tag{8}$$

where Φ_i is the i'th activation layer of VGG-19 pre-trained network on ImageNet [57].

Style loss is shown by Sajjadi et al. [58] as an effective way to deal with "checkerboard" artifacts caused by transpose convolution [59]. Style loss adopts the same activation layers as perceptual loss and is defined as:

$$L_s = \mathbb{E}_j\left[\|G_j^{\Phi_i}(\mathbf{I}_p) - G_j^{\Phi_i}(\mathbf{I}_{GT})\|_1\right], \tag{9}$$

where $G_j^{\Phi_i}$ is the Gram matrix of activation map Φ_i. Total variation loss [60] is used for smoothing the output spatially and compacting the possible noise in the decoder. Total variation loss for an $H \times W \times C$ feature map is defined as:

$$L_{TV} = \frac{1}{HWC} \sum_{i,j,k} \sqrt{(\mathbf{I}_{i,j+1,k} - \mathbf{I}_{i,j,k})^2 + (\mathbf{I}_{i+1,j,k} - \mathbf{I}_{i,j,k})^2}. \tag{10}$$

The final training loss for ITrans network is:

$$\begin{aligned} L_{inpaint} = &\ \lambda_{l1}L_{l1} + \lambda_{advi}L_{advi} + \lambda_p L_p \\ &+ \lambda_s L_s + \lambda_{TV} L_{TV}. \end{aligned} \tag{11}$$

In the training settings, we set $\lambda_{l1} = 1$, $\lambda_{advi} = 0.1$, $\lambda_p = 0.1$, $\lambda_s = 250$, and $\lambda_{TV} = 0.01$.

4 Experiments

4.1 Implementation details

The ITrans network is implemented in PyTorch [62]. We use Adam [63] optimizer with β_1=0 and β_2=0.9. The learning rate of the generator is set to 10^{-4} learning rate initially, and decreases to 10^{-5} until convergence. The discriminator's learning rate is one-tenth that of the generator. In the edge

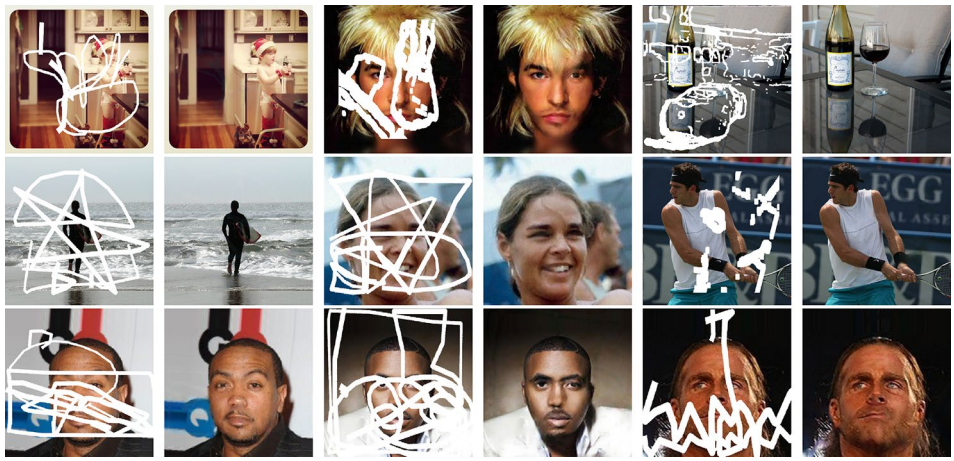

Fig. 1 Inpainting results by the proposed ITrans network for diverse scenes and human faces

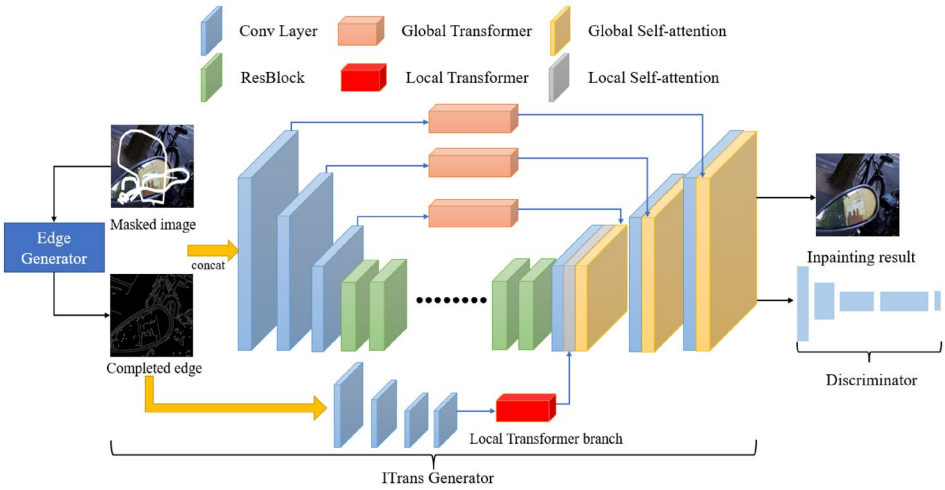

Fig. 2 ITrans Network. In the ITrans network, we adopt an encoder–decoder structure along with ResBlock bottleneck. The global transformer is added as skip layer to gather encoder and decoder features and self-attention together. We add another branch specifically for the local transformer, which aims at extracting fine image details

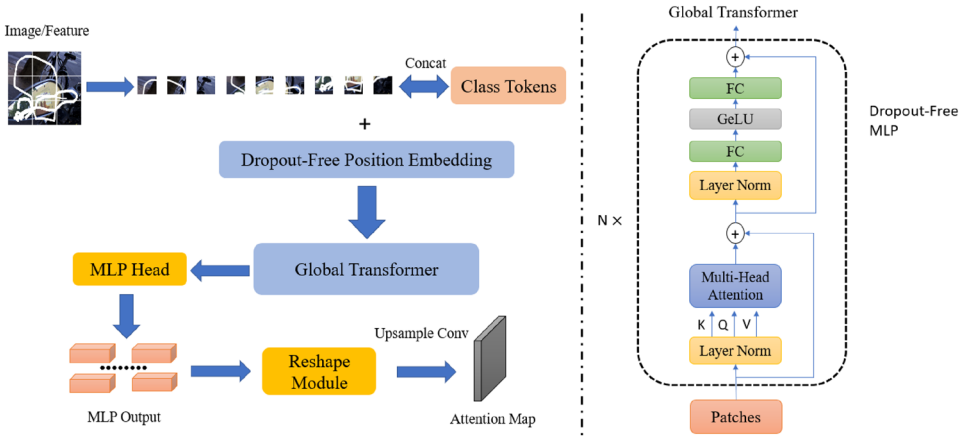

Fig. 3 Global transformer. In global transformer, we do not adopt dropout layer in order to keep all pixels in the feature map. We believe it is essential to hold all pixels in our attention map

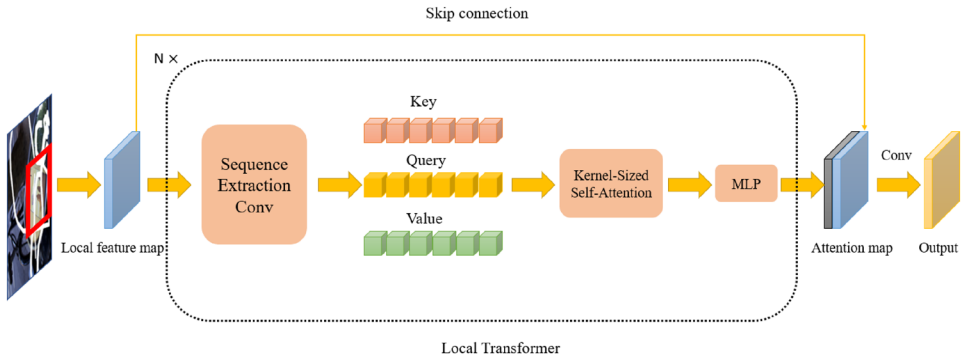

Fig. 4 Local transformer. We obtain transformer sequences with sequence extraction convolution. Self-attention is computed by a kernel-sized attention layer. And we add a skip layer at the output stage to combine input and self-attention together for feature aggregating

completion stage, the initial edge is generated by the Canny edge detector [45].

4.2 Training datasets

Our ITrans network is trained on the MS-COCO [64], Places2 [65] datasets for places inpainting and CelebA [66] dataset for human faces. Places2 is an image inpainting dataset that contains over 8 million images with more than 365 places categories, while CelebA has over 200 thousand celebrity faces. To improve free-form inpainting performance, we mix NVIDIA-ALDR datasets [8] and Google-Quick-Draw!-based QD-IMD [67] together, as well as randomly produced square masks. Both of these datasets include randomly drawn stripes to simulate the artifacts present in real-world inpainting tasks. The resolution of training images is set to 512×512 and all models are trained for 1 million iterations with a batchsize of 8.

5 Results

In our experiments, we use Places2 and CelebA for places and human faces tests respectively, and NVIDIA-ALDR test sets are used for different mask regions.

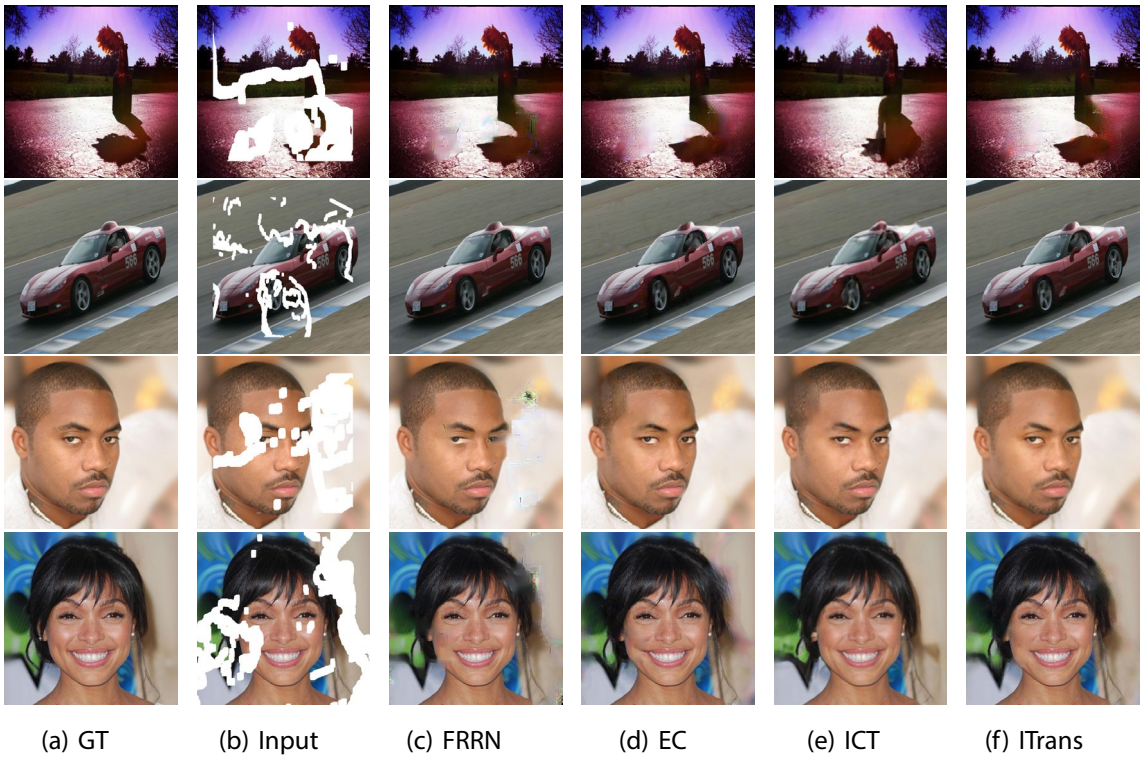

Fig. 5 Qualitative comparison with current models. **a** Ground truth, **b** masked images, **c** FRRN [61], **d** EdgeConnect [9], **e** ICT [43], **f** ITrans

Table 1 Quantitative results on Places2

Mask ratio		FRRN	EC	ICT	Itrans
PSNR	0–10%	31.33	31.92	26.60	**32.15**
	10–20%	27.73	27.60	24.11	**27.88**
	20–30%	24.53	24.70	21.98	**24.98**
	30–40%	22.16	22.50	20.25	**22.79**
	40–50%	20.06	20.69	18.83	**21.05**
	All	23.83	24.36	21.66	**24.68**
SSIM	0–10%	0.9677	0.9712	0.9191	**0.9726**
	10–20%	0.9358	0.9342	0.8719	**0.9375**
	20–30%	0.8795	0.8808	0.8090	**0.8868**
	30–40%	0.8077	0.8132	0.7369	**0.8222**
	40–50%	0.7199	0.7324	0.6563	**0.7456**
	All	0.8113	0.8203	0.7582	**0.8288**
MAE(%)	0–10%	1.78	1.56	3.37	**1.54**
	10–20%	2.22	2.28	4.17	**2.22**
	20–30%	3.23	3.26	5.25	**3.14**
	30–40%	4.51	4.46	6.51	**4.25**
	40–50%	6.19	5.91	7.98	**5.61**
	All	4.86	4.36	6.14	**4.17**
FID	0–10%	4.49	4.08	15.03	**4.06**
	10–20%	11.52	10.69	20.14	**10.31**
	20–30%	22.94	20.43	26.52	**19.14**
	30–40%	38.04	34.77	33.67	**31.84**
	40–50%	58.73	53.94	**43.99**	49.02
	All	17.10	16.24	**14.37**	15.10

We compare our Inpainting Transformer network (ITrans) with FRRN [61], EdgeConnect [9] and ICT [43]. On evaluating metrics, PSNR and SSIM are higher the better, while MAE and FID are lower the better. The best metrics are boldfaced

Table 2 Quantitative results on Celeb-HQ

Mask ratio		FRRN	EC	ICT	Itrans
PSNR	0–10%	35.23	35.29	32.45	**35.67**
	10–20%	30.72	30.89	29.78	**31.19**
	20–30%	27.56	27.73	27.34	**27.97**
	30–40%	24.86	24.90	25.15	**25.28**
	40–50%	22.16	22.75	**23.50**	23.30
	All	26.18	26.70	26.55	**27.28**
SSIM	0–10%	0.9804	0.9802	0.9674	**0.9815**
	10–20%	0.9558	0.9563	0.9462	**0.9584**
	20–30%	0.9182	0.9185	0.9153	**0.9214**
	30–40%	0.8640	0.8662	0.8711	**0.8742**
	40–50%	0.7901	0.8007	**0.8283**	0.8098
	All	0.8547	0.8629	0.8704	**0.8791**
MAE(%)	0–10%	1.09	1.09	1.63	**1.06**
	10–20%	1.58	1.57	2.03	**1.50**
	20–30%	2.26	2.27	2.60	**2.14**
	30–40%	3.28	3.21	3.34	**3.00**
	40–50%	4.82	4.44	4.16	**4.03**
	All	3.84	3.32	3.23	**2.99**
FID	0–10%	3.32	3.21	15.19	**2.61**
	10–20%	8.67	7.77	17.87	**7.04**
	20–30%	15.97	14.74	21.16	**14.16**
	30–40%	25.14	25.06	25.02	**23.34**
	40–50%	37.87	36.46	**27.98**	34.79
	All	14.17	12.56	14.07	**11.04**

The best metrics are boldfaced

5.1 Qualitative comparison

Figure 1 shows inpainting results obtained by our ITrans network. Our ITrans network produces visually realistic results when the missing area is extensive. In Fig. 5, we compare images generated by our model to those generated by other inpainting approaches. ITrans works well on fine details, demonstrating the efficacy of our network structure. With the use of edge maps, ITrans network could specifically concentrate on pixel generation with the transformer-based self-attention.

5.2 Quantitative comparison

We use four quantitative metrics to evaluate inpainting qualities: (1) relative L1 (MAE); (2) structural similarity index (SSIM) [68]; (3) peak signal-to-noise ratio (PSNR); (4) Frechet inception distance (FID) [69].

Table 3 Inpainting results with and without global transformer

Dataset	Places		CelebA	
Mask	w/o G-Trans	w/ G-Trans	w/o G-Trans	w/ G-Trans
PSNR				
20–30%	24.70	24.73	27.73	27.79
30–40%	22.50	22.53	24.86	25.10
SSIM				
20–30%	0.8808	0.8824	0.9185	0.9211
30–40%	0.8132	0.8150	0.8640	0.8704
MAE				
20–30%	3.26	3.23	2.27	2.21
30–40%	4.46	4.40	3.28	3.14
FID				
20–30%	20.43	20.23	14.74	14.66
30–40%	34.77	34.73	25.06	25.00

Statistics are based on 2000 images on Places2 and 500 images on Celeb-HQ, respectively

Pixel-wise metrics measures the accuracy (MAE), structure (SSIM) and color (PSNR) of inpainting images with ground truths. FID measures perceptually accuracy due to its feature-based characteristic, which is based on the Inception-V3 model [70] for superior perception performance than humans [71–73].

Table 1 shows our experimental results on Places2, and Table 2 shows our testing results on Celeb-HQ. The Places2 dataset includes 12,000 images, with each mask ratio consisting of 2000 masks. Celeb-HQ comprises 500 images for each mask ratio and 2000 images for all mask regions. We compare the ITrans network with FRRN [61], EdgeConnect [9] and ICT [43]. We obtain statistics using available codes and pretrained weights. Our experiments demonstrate that our ITrans network exceeds other approaches on the majority of metrics. However, it should be noted that ICT outperforms better than ITrans in terms of large masks, especially on human faces. We believe that this is because visual plausibility is more essential than restoring the original images in large masks.

5.3 Ablation study

In this section, we will turn to our key contributions: the global and local transformer. We will demonstrate their efficacy through the following ablation studies.

Global transformer Skip layers are widely used in the encoder–decoder structure. However, traditional skip layers simply combine the encoder and decoder without any extra structure. In contrast, our global transformer aggregates encoder attention and decoder features in the skip layers. Table 3 shows the inpainting performance with and without global transformer. The results reveal that our global transformer outperforms the network without a skip layer. This suggests that our global transformer performs well on inpainting tasks and demonstrates the efficacy of global attention.

Table 4 Inpainting results with and without local transformer

Dataset	Places		CelebA	
Mask	w/o L-Trans	w/ L-Trans	w/o L-Trans	w/ L-Trans
PSNR				
20–30%	24.73	24.97	27.79	27.88
30–40%	22.53	22.76	25.10	25.26
SSIM				
20–30%	0.8824	0.8858	0.9211	0.9214
30–40%	0.8150	0.8204	0.8704	0.8735
MAE				
20–30%	3.23	3.17	2.21	2.14
30–40%	4.40	4.28	3.14	3.02
FID				
20–30%	20.23	19.57	14.66	14.16
30–40%	34.73	32.57	25.00	23.56

Statistics are based on 2000 images on Places2 and 500 images on Celeb-HQ, respectively

Fig. 6 Failure cases. Artifacts appear in huge missing holes

Local transformer The local transformer is the next focus of our research in the ITrans network, as it contains both local and global transformers. Having already demonstrated the effectiveness of the global transformer, we are now gaining experience with the local transformer. We compare the performance of the network with and without the local transformer, and the results are shown in Table 4. Our findings demonstrate that our proposed local transformer module effectively enhances inpainting performance for both places and faces, with the additional branch of local attention proving highly valuable. This additional self-attention branch highlights the importance of detailed local self-attention in improving the network's inpainting ability.

5.4 Limitations

Failure cases are shown in Fig. 6. Blurriness and artifacts appear when the inpainting mask is large or complicated. A better edge completion model and a better network structure, we believe, might improve performance. Moreover, the current generation performance of transformers is relatively poor, we need to discover a solution to enhance their generating ability. Even though a 512×512 image is sufficient, our model still need to be experimented on higher resolution to enhance the utility of our ITrans network.

5.5 Future works

The current ITrans network has significant scope for improvement. For example, the network needs to be trained on a wider variety of datasets. Although the current dataset provides simulation of various contexts, it is still insufficient for real-world inpainting tasks. Additionally, the transformer structure remains computationally expensive during training and requires a lighter version. Recently, diffusion models ?? have become popular in generative tasks, and a combination

of CNNs, transformers, and diffusion models could hold great promise in this field.

6 Conclusion

Through multiple experiments, we have evaluated the end-to-end ITrans network's ability to perform well in various inpainting scenarios. The ITrans network leverages the inductive bias of CNNs while adding flexibility with its global and local transformers. The global transformer provides global semantic self-attention for encoder feature maps, which are then utilized in the decoder. The local transformer extracts local feature details to enhance the inpainting results further. Finally, future enhancements of the generating ability are expected to improve overall performance.

Author contributions All authors reviewed the manuscript. WM wrote the main manuscript text, performed all experiments and prepared all figures. LW wrote the abstract and rounded off the manuscript text. HL provided experiment equipment.

Funding Open Access funding provided by University of Jyväskylä (JYU).

Data availability The data mentioned in this article is free to use.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Criminisi, A., Perez, P., Toyama, K.: Object removal by exemplar-based inpainting. In: 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings., vol. 2, p. 2003. IEEE
2. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
3. Liang, J., Zeng, H., Cui, M., Xie, X., Zhang, L.: Ppr10k: a large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 653–661 (2021)
4. Hays, J., Efros, A.A.: Scene completion using millions of photographs. ACM Trans. Graph. (ToG) **26**(3), 4 (2007)
5. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: a randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
6. 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, pp. 2536–2544 (2016)
7. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Globally and locally consistent image completion. ACM Trans. Graph. (ToG) **36**(4), 1–14 (2017)
8. 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), pp. 85–100 (2018)
9. Nazeri, K., Ng, E., Joseph, T., Qureshi, F., Ebrahimi, M.: EdgeConnect: structure guided image inpainting using edge prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 0–0 (2019)
10. Xiong, W., Yu, J., Lin, Z., Yang, J., Lu, X., Barnes, C., Luo, J.: Foreground-aware image inpainting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (ICCV), pp. 5840–5848 (2019)
11. 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), pp. 4471–4480 (2019)
12. Song, Y., Yang, C., Shen, Y., Wang, P., Huang, Q., Kuo, C.-C.J.: SPG-Net: segmentation prediction and guidance network for image inpainting. arXiv preprint arXiv:1805.03356 (2018)
13. Zeng, Y., Lin, Z., Yang, J., Zhang, J., Shechtman, E., Lu, H.: High-resolution image inpainting with iterative confidence feedback and guided upsampling. In: European Conference on Computer Vision, pp. 1–17. Springer (2020)
14. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7508–7517 (2020)
15. Zhao, S., Cui, J., Sheng, Y., Dong, Y., Liang, X., Chang, E.I., Xu, Y.: Large scale image completion via co-modulated generative adversarial networks. arXiv preprint arXiv:2103.10428 (2021)
16. 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, pp. 9371–9381 (2021)
17. Liu, R., Wang, X., Lu, H., Wu, Z., Fan, Q., Li, S., Jin, X.: SCCGAN: style and characters inpainting based on CGAN. Mob. Netw. Appl. **26**(1), 3–12 (2021)
18. Wang, L., Zhang, S., Gu, L., Zhang, J., Zhai, X., Sha, X., Chang, S.: Automatic consecutive context perceived transformer GAN for serial sectioning image blind inpainting. Comput. Biol. Med. **136**, 104751 (2021)
19. 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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5741–5750 (2020)
20. d'Ascoli, S., Touvron, H., Leavitt, M.L., Morcos, A.S., Biroli, G., Sagun, L.: ConViT: improving vision transformers with soft convolutional inductive biases. In: International Conference on Machine Learning, pp. 2286–2296. PMLR (2021)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

22. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
23. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2149–2159 (2022)
24. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241. Springer (2015)
25. 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, pp. 5505–5514 (2018)
26. Kim, Y., Cheon, M., Lee, J.: Texture transform attention for realistic image inpainting. arXiv preprint arXiv:2012.04242 (2020)
27. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
28. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
29. 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, pp. 770–778 (2016)
30. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, pp. 10347–10357. PMLR (2021)
31. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229. Springer (2020)
32. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)
33. Wang, H., Zhu, Y., Adam, H., Yuille, A., Chen, L.-C.: Max-DeepLab: end-to-end panoptic segmentation with mask transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5463–5474 (2021)
34. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8741–8750 (2021)
35. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6881–6890 (2021)
36. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., Tran, D.: Image transformer. In: International Conference on Machine Learning, pp. 4055–4064. PMLR (2018)
37. Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., Sutskever, I.: Generative pretraining from pixels. In: International Conference on Machine Learning, pp. 1691–1703. PMLR (2020)
38. Jiang, Y., Chang, S., Wang, Z.: TransGAN: two transformers can make one strong GAN. arXiv preprint arXiv:2102.07074 (2021)
39. Lee, K., Chang, H., Jiang, L., Zhang, H., Tu, Z., Liu, C.: ViTGAN: training GANs with vision transformers. arXiv preprint arXiv:2107.04589 (2021)
40. Kim, H., Papamakarios, G., Mnih, A.: The Lipschitz constant of self-attention. In: International Conference on Machine Learning, pp. 5562–5571. PMLR (2021)
41. Yu, Y., et al.: Diverse image inpainting with bidirectional and autoregressive transformers. arXiv (2021)
42. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)
43. 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, pp. 4692–4701 (2021)
44. Zheng, C., Cham, T.-J., Cai, J., Phung, D.: Bridging global context interactions for high-fidelity image completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11512–11522 (2022)
45. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 679–698 (1986)
46. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27** (2014)
47. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
48. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018)
49. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Improved texture networks: maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6924–6932 (2017)
50. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
51. Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., et al.: An image is worth 16x16 words: transformers for image recognition at scale (2021)
52. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
53. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018)
54. 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 (2016). Springer
55. 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, pp. 2414–2423 (2016)
56. Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. Adv. Neural. Inf. Process. Syst. **28**, 262–270 (2015)
57. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
58. Sajjadi, M.S., Scholkopf, B., Hirsch, M.: EnhanceNet: single image super-resolution through automated texture synthesis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4491–4500 (2017)
59. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distillation **1**(10), 3 (2016)

60. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D **60**(1–4), 259–268 (1992)
61. Guo, Z., Chen, Z., Yu, T., Chen, J., Liu, S.: Progressive image inpainting with full-resolution residual network. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2496–2504 (2019)
62. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: an imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. **32** (2019)
63. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. ICLR **9** (2015). arXiv preprint arXiv:1412.6980
64. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
65. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: a 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)
66. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
67. Iskakov, K.: Semi-parametric image inpainting. arXiv preprint arXiv:1807.02855 (2018)
68. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
69. 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. Adv. Neural Inf. Process. Syst. **30** (2017)
70. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
71. Dolhansky, B., Ferrer, C.C.: Eye in-painting with exemplar generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7902–7911 (2018)
72. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International Conference on Machine Learning, pp. 7354–7363. PMLR (2019)
73. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.