High-Fidelity Pluralistic Image Completion with Transformers

Ziyu Wan[1]   Jingbo Zhang[1]   Dongdong Chen[2]   Jing Liao[1*]

[1]City University of Hong Kong   [2]Microsoft Cloud + AI
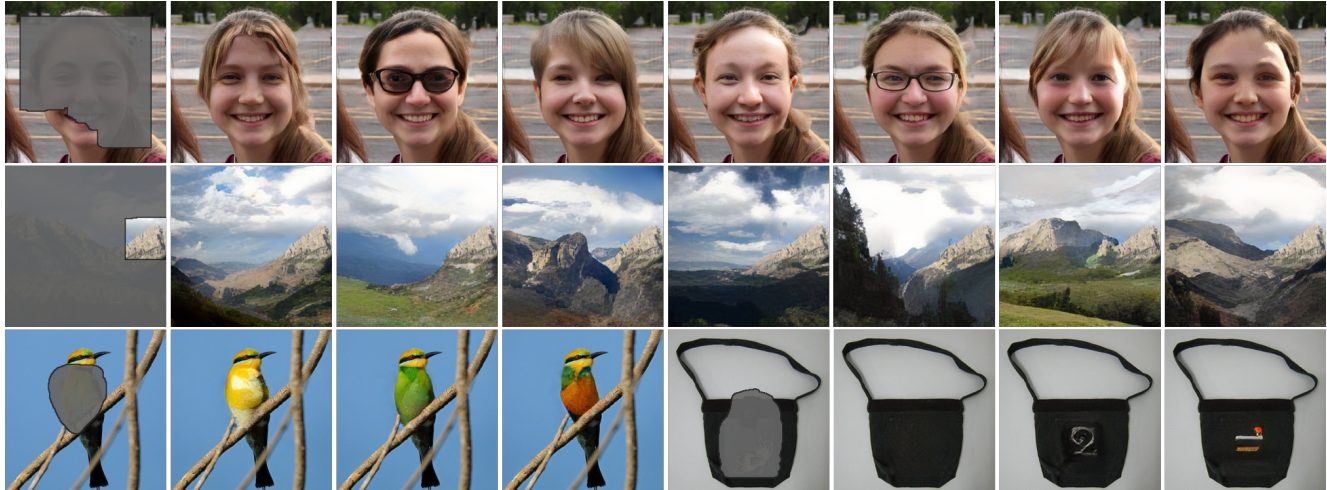http://raywzy.com/ICT

Figure 1: Pluralistic free-form image completion results produced by our method.

Abstract

Image completion has made tremendous progress with convolutional neural networks (CNNs), because of their powerful texture modeling capacity. However, due to some inherent properties (e.g., local inductive prior, spatial-invariant kernels), CNNs do not perform well in understanding global structures or naturally support pluralistic completion. Recently, transformers demonstrate their power in modeling the long-term relationship and generating diverse results, but their computation complexity is quadratic to input length, thus hampering the application in processing high-resolution images. This paper brings the best of both worlds to pluralistic image completion: appearance prior reconstruction with transformer and texture replenishment with CNN. The former transformer recovers pluralistic coherent structures together with some coarse textures, while the latter CNN enhances the local texture details of coarse priors guided by the high-resolution masked images. The proposed method vastly outperforms state-of-the-art methods in terms of three aspects: 1) large performance boost on image fidelity even compared to deterministic completion methods; 2) better diversity and higher fidelity for pluralistic completion; 3) exceptional general-

ization ability on large masks and generic dataset, like ImageNet. Code and pre-trained models have been publicly released at https://github.com/raywzy/ICT.

1. Introduction

Image completion (a.k.a. image inpainting), which aims to fill the missing parts of images with visually realistic and semantically appropriate contents, has been a longstanding and critical problem in computer vision areas. It is widely used in a broad range of applications, such as object removal [2], photo restoration [34, 35], image manipulation [23, 15], and image re-targeting [6]. To solve this challenging task, traditional methods like PatchMatch [2] usually search for similar patches within the image and paste them into the missing regions, but they require appropriate information to be contained in the input image, *e.g.*, similar structures or patches, which is often difficult to satisfy.

In recent years, CNN-based solutions [22, 14, 18, 27, 36, 21] started to dominate this field. By training on large-scale datasets in a self-supervised way, CNNs have shown their strength in learning rich texture patterns, and fills the missing regions with such learned patterns. Moreover, CNN models are computationally efficient considering the sparse connectivity of convolutions. Nonetheless, they share some

---

*Corresponding author.

inherent limitations: 1) The local inductive priors of convolution operation make modeling the global structures of an image difficult; 2) CNN filters are spatial-invariant, *i.e.* the same convolution kernel operates on the features across all positions, by that the duplicated patterns or blurry artifacts frequently appear in the masked regions. On the other hand, CNN models are inherently deterministic. To achieve diverse completion outputs, recent frameworks [41, 39] rely on optimizing the variational lower bound of instance likelihood. Nonetheless, extra distribution assumption would inevitably hurt the quality of generated contents [40].

Transformer, as well-explored architectures in language tasks, is on-the-rise in many computer vision tasks. Compared to CNN models, it abandons the baked-in local inductive prior and is designed to support the long-term interaction via the dense attention module [33]. Some preliminary works [5] also demonstrate its capacity in modeling the structural relationships for natural image synthesis. Another advantage of using a transformer for synthesis is that it naturally supports pluralistic outputs by directly optimizing the underlying data distribution. However, the transformer also has its own deficiency. Due to quadratically increased computational complexity with input length, it struggles in high-resolution image synthesis or processing. Moreover, most existing transformer-based generative models [26, 5] works in an auto-regressive manner, *i.e.*, synthesize pixels in a fixed order, like the raster-scan order, thus hampering its application in the image completion task where the missing regions are often with arbitrary shapes and sizes.

In this paper, we propose a new high-fidelity pluralistic image completion method by bringing the best of both worlds: the global structural understanding ability of transformers and the efficient local texture refinement ability of CNN models. To achieve this, we decouple image completion into two steps: pluralistic appearance priors reconstruction with a transformer to recover the coherent image structures, and low-resolution upsampling with CNN to replenish fine textures. Specifically, given an input image with missing regions, we first leverage the transformer to sample low-resolution completion results, *i.e.* appearance priors. Then, guided by the appearance priors and the available pixels of the input image, another upsampling CNN model is utilized to render high-fidelity textures for missing regions meanwhile ensuring coherence with neighboring pixels. In particular, unlike previous auto-regressive methods [5, 32], in order to make the transformer model capable of completing the missing regions by considering all the available contexts, we optimize the log-likelihood objective of missing pixels based on the bi-directional conditions, which is inspired by the masked language model like BERT [9].

To demonstrate the superiority, we compare our method with state-of-the-art deterministic [25, 20, 37] and pluralistic [41] image completion approaches on multiple datasets.

Our method makes significant progress from three aspects: 1) Compared with previous deterministic completion methods, our method establishes a new state of the art and outperforms theirs on a variety of metrics by a large margin; 2) Compared with previous pluralistic completion methods, our method further enhances the results diversity, meanwhile achieving higher completion fidelity; 3) Thanks to the strong structure modeling capacity of transformers, our method generalizes much better in completing extremely large missing region and large-scale generic datasets (*e.g.*, ImageNet) as shown in Figure. 1. Remarkably, the FID score on ImageNet is improved by 41.2 at most compared with the state-of-the-art method PIC [41].

2. Related Works

Visual Transformers Vaswani *et al.* [33] firstly propose transformers for machine translation, whose success subsequently has been proved in various down-stream natural language processing (NLP) tasks. The overall architecture of transformers is composed of stacked self-attention and point-wise feed-forward layers for encoder and decoder. Since the attention mechanism could model the dense relationship among elements of input sequence well, transformers are now gradually popular in computer vision areas. For example, DETR [4] employ transformers as the backbone to solve the object detection problem. Dosovitskiy *et al.* [10] propose ViT, which firstly utilize transformers in the image recognition area and achieved excellent results compared with CNN-based methods. Moreover, Parmar *et al.* [26] and Chen *et al.* [5] leverage the transformer to model the image. Nonetheless, to generate an image, these methods rely on a fixed permutation order, which is not suitable to complete the missing areas with varying shapes.

Deterministic Image Completion Traditional image completion methods, like diffusion-based [3, 11] and patch-based [2, 12, 7], rely on strong low-level assumptions, which may be violated while facing large-area masks. To generate semantic-coherent contents, recently many CNN-based methods [27, 19, 21, 36, 25] have been proposed. Most of the methods share a similar encoder-decoder architecture. Specifically, Pathak *et al.* [27] bring the adversarial training into inpainting and achieve semantic hole-filling. Iizuka *et al.* [14] improve the performance of CE [27] by involving a local-global discriminator. Yu *et al.* [36] propose a new contextual attention module to capture the long-range correlations. Liu *et al.* [19] design a new operator named partial-conv to alleviate the negative influence of the masked regions produced by convolutions. These methods could generate reasonable contents for masked regions but lack the ability to generate diversified results.

Pluralistic Image Completion To obtain a diverse set of results for each masked input, Zheng *et al.* [41] propose a dual pipeline framework, which couples the esti-
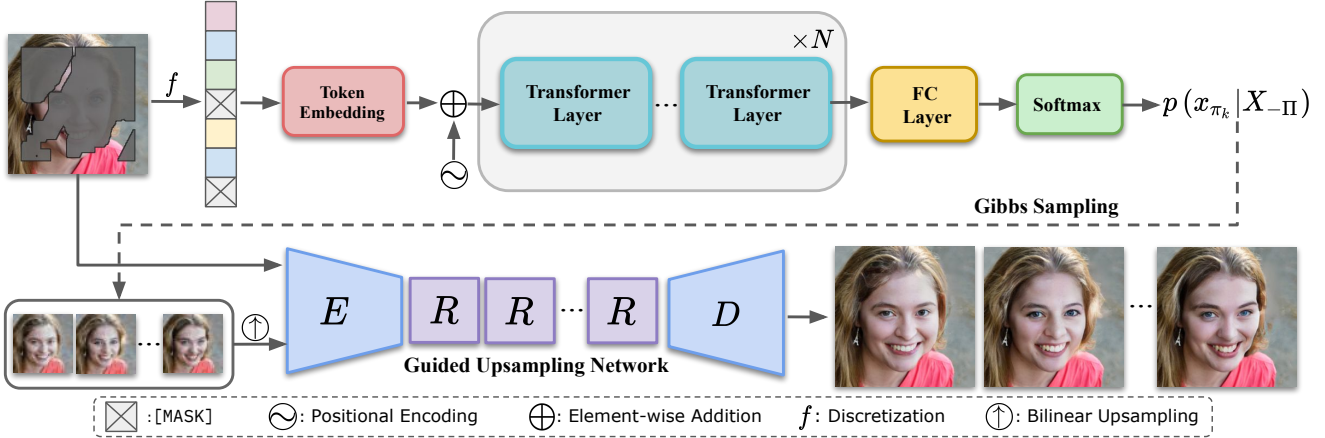
Figure 2: **Pipeline Overview.** Our method consists of two networks. The above one is **bi-directional transformer**, which is responsible for producing the probability distribution of missing regions, then the appearance priors could be reconstructed by sampling from this distribution with diversities. Subsequently, we employ another **CNN** to upsample the appearance prior to original resolution under the guidance of input masked images. Our method combines both advantages of transformer and CNN, leading to high-fidelity pluralistic image completion performance. **E**: Encoder, **D**: Decoder, **R**: Residual block.

mated distribution from the reconstructive path and conditional prior of the generative path via jointly maximizing the lower bound. Similar with [41], UCTGAN [39] project both the masked input and reference image into a common space via optimizing the KL-divergence between encoded features and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ distribution to achieve diversified sampling. Although they have achieved some diversities to a certain extent, their completion qualities are limited due to variational training. Unlike these methods, we directly optimize the log-likelihood in the discrete space via transformers without auxiliary assumptions.

3. Method

Image completion aims to transform the input image $I_m \in \mathbb{R}^{H \times W \times 3}$ with missing pixels into a complete image $I \in \mathbb{R}^{H \times W \times 3}$. This task is inherently stochastic, which means given the masked image I_m, there exists a conditional distribution $p(I|I_m)$. Since obtaining coarse prior X given I and I_m is deterministic, then $p(I|I_m)$ could be rewritten as,

$$
\begin{aligned}
p(I|I_m) &= p(I|I_m) \cdot p(X|I, I_m) \\
&= p(X, I|I_m) \\
&= p(X|I_m) \cdot p(I|X, I_m).
\end{aligned}
\tag{1}
$$

Instead of directly sampling from $p(I|I_m)$, we first use a transformer to model the underlying distribution of appearance priors given I_m, denoted as $p(X|I_m)$ (described in Sec. 3.1). These reconstructed appearance priors contain ample cues of global structure and coarse textures, thanks to the transformer's strong representation ability. Subsequently, we employ another CNN to replenish texture details under the guidance of appearance priors and unmasked

pixels, denoted as $p(I|X, I_m)$ (described in Sec. 3.2). The overall pipeline could be found in Figure. 2.

3.1. Appearance Priors Reconstruction

Discretization Considering the quadratically increasing computational cost of multi-head attention [33] in the transformer architecture, we represent the appearance priors of a natural image with its corresponding low-resolution version (32×32 or 48×48 in our implementation), which contains structural information and coarse textures only. Nonetheless, the dimensionality of RGB pixel (256^3) is still too large. To further reduce the dimension and faithfully represent the low-resolution image, an extra visual vocabulary with spatial size 512×3 is generated using K-Means cluster centers of the whole ImageNet [8] pixel spaces. Then for each pixel of appearance priors, we search the index of the nearest element from the visual vocabulary to obtain its discrete representation. In addition, the elements of the representation sequence corresponding to hole regions will be replaced with a special token [MASK], which is also the learning target of the transformer. To this end, we convert the appearance prior into a discretized sequence.

Transformer For each token of the discretized sequence $X = \{x_1, x_2, \cdots, x_\mathbb{L}\}$, where $\mathbb{L}$ is the length of X, we project it into a $d-$dimensional feature vector through prepending a learnable embedding. To encode the spatial information, extra learnable position embeddings will be added into the token features for every location $1 \leq i \leq \mathbb{L}$ to form the final input $E \in \mathbb{R}^{\mathbb{L} \times d}$ of transformer model.

Following GPT-2 [28], we use decoder-only transformer as our network architecture, which is mainly composed with N self-attention based transformer layers. At each trans-
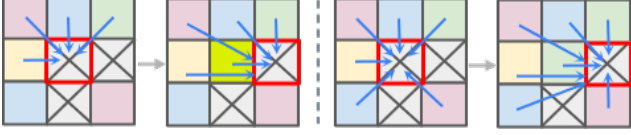
Figure 3: Differences between single-directional (left) and bi-directional (right) attention.

former layer l, the calculation could be formulated as

$$F^{l-1} = \text{LN}(\text{MSA}(E^{l-1})) + E^{l-1}$$
$$E^l = \text{LN}(\text{MLP}(F^{l-1})) + F^{l-1}, \tag{2}$$

where LN, MSA, MLP denote layer normalization [1], multi-head self-attention and FC layer respectively. More specifically, given input E, the MSA could be computed as:

$$head_j = \text{softmax}\left(\frac{E\mathbf{W}_Q^j (E\mathbf{W}_K^j)^T}{\sqrt{d}}\right)(E\mathbf{W}_V^j)$$
$$\text{MSA} = [head_1; ...; head_h]\mathbf{W}_O, \tag{3}$$

where h is the number of head, $\mathbf{W}_Q^j$, $\mathbf{W}_K^j$ and $\mathbf{W}_V^j$ are three learnable linear projection layers, $1 \le j \le h$. $\mathbf{W}_O$ is also a learnable FC layer, whose target is to fuse the concatenation of the outputs from different heads. By adjusting the parameters of transformer layer N, embedding dimension d and head number h, we could easily scale the size of the transformer. It should also be noted that unlike autoregressive transformers [5, 26], which generate elements via single-directional attention, *i.e.* only constrained by the context before scanning line, we make each token attend to all positions to achieve bi-directional attention, as shown in Figure. 3, which effectively maintains the semantic consistency between generated contents and unmasked regions.

The output of the final transformer layer is further projected to a per-element distribution over 512 elements of visual vocabulary with the fully connected layers and softmax function. We adopt the masked language model (MLM) objective similar as the one used in BERT [9] to optimize the transformer model. Specifically, let $\Pi = \{\pi_1, \pi_2, \ldots, \pi_K\}$ denote the indexes of [MASK] tokens in the discretized input, where K is the number of masked tokens. Let X_Π denote the set of [MASK] tokens in X, and $X_{-\Pi}$ denote the set of unmasked tokens. The objective of MLM minimizes the negative log-likelihood of X_Π conditioned on all observed regions:

$$L_{MLM} = \mathbb{E}_X\left[\frac{1}{K}\sum_{k=1}^{K} -\log p(x_{\pi_k}|X_{-\Pi}, \theta)\right], \tag{4}$$

where θ is the parameters of transformer. MLM objective incorporating with bi-directional attention ensures that the transformer model could capture the whole contextual information to predict the probability distribution of missing regions.

Sampling Strategy We introduce how to obtain reasonable and diversified appearance priors using the trained transformer in this section. Given the generated distribution of the transformer, directly sampling the entire set of masked positions does not produce plausible results due to the independence property. Instead, we employ Gibbs sampling to iteratively sample tokens at different locations. Specifically, in each iteration, a grid position is sampled from $p(x_{\pi_k}|X_{-\Pi}, X_{<\pi_k}, \theta)$ with the top-$\mathcal{K}$ predicted elements, where $X_{<\pi_k}$ denotes the previous generated tokens. Then the corresponding [MASK] token is replaced with the sampled one, and the process is repeated until all positions are updated. Similar with PixelCNN [32], the positions are sequentially chosen in a raster-scan manner by default. After sampling, we could obtain a bunch of complete token sequences. For each complete discrete sequence sampled from transformer, we reconstruct its appearance priors $X \in \mathbb{R}^{\mathbb{L}\times 3}$ with querying the visual vocabulary.

3.2. Guided Upsampling

After reconstructing the low-dimensional appearance priors, we reshape X into $I_t \in \mathbb{R}^{\sqrt{\mathbb{L}}\times\sqrt{\mathbb{L}}\times 3}$ for subsequent processing. Since I_t has contained the diversity, now the problem is how to learn a deterministic mapping to re-scale the I_t into original resolution $H \times W \times 3$, meanwhile preserving the boundary consistency between hole regions and unmasked regions. To achieve this point, since CNNs have advantages of modeling texture patterns, here we introduce another guided upsampling network, which could render high-fidelity details of reconstructed appearance priors with the guidance of masked input I_m. The processing of guided upsampling could be written as

$$I_{pred} = \mathcal{F}_\delta(I_t^\uparrow \frown I_m) \in \mathbb{R}^{H\times W\times 3}, \tag{5}$$

where $I_t^\uparrow$ is the result of bilinear interpolation of I_t and $\frown$ denotes the concatenation operation along the channel dimension. $\mathcal{F}$ is the backbone of upsampling network parameterized by δ, which is mainly composed of encoder, decoder and several residual blocks. More details about the architecture could be found in the supplementary material.

We optimize this guided upsampling network by minimizing ℓ_1 loss between I_{pred} and corresponding ground-truth I:

$$L_{\ell_1} = \mathbb{E}(\|I_{pred} - I\|_1). \tag{6}$$

To generate more realistic details, extra adversarial loss is also involved in the training process, specifically,

$$L_{adv} = \mathbb{E}\left[\log(1 - \mathcal{D}_\omega(I_{pred}))\right] + \mathbb{E}\left[\log \mathcal{D}_\omega(I)\right], \tag{7}$$

where $\mathcal{D}$ is the discriminator parameterized by ω. We jointly train upsampling network $\mathcal{F}$ and discriminator $\mathcal{D}$ through solving the following optimization,

$$\min_{\mathcal{F}}\max_{\mathcal{D}} L_{upsample}(\delta, \omega) = \alpha_1 L_{\ell_1} + \alpha_2 L_{adv}. \tag{8}$$
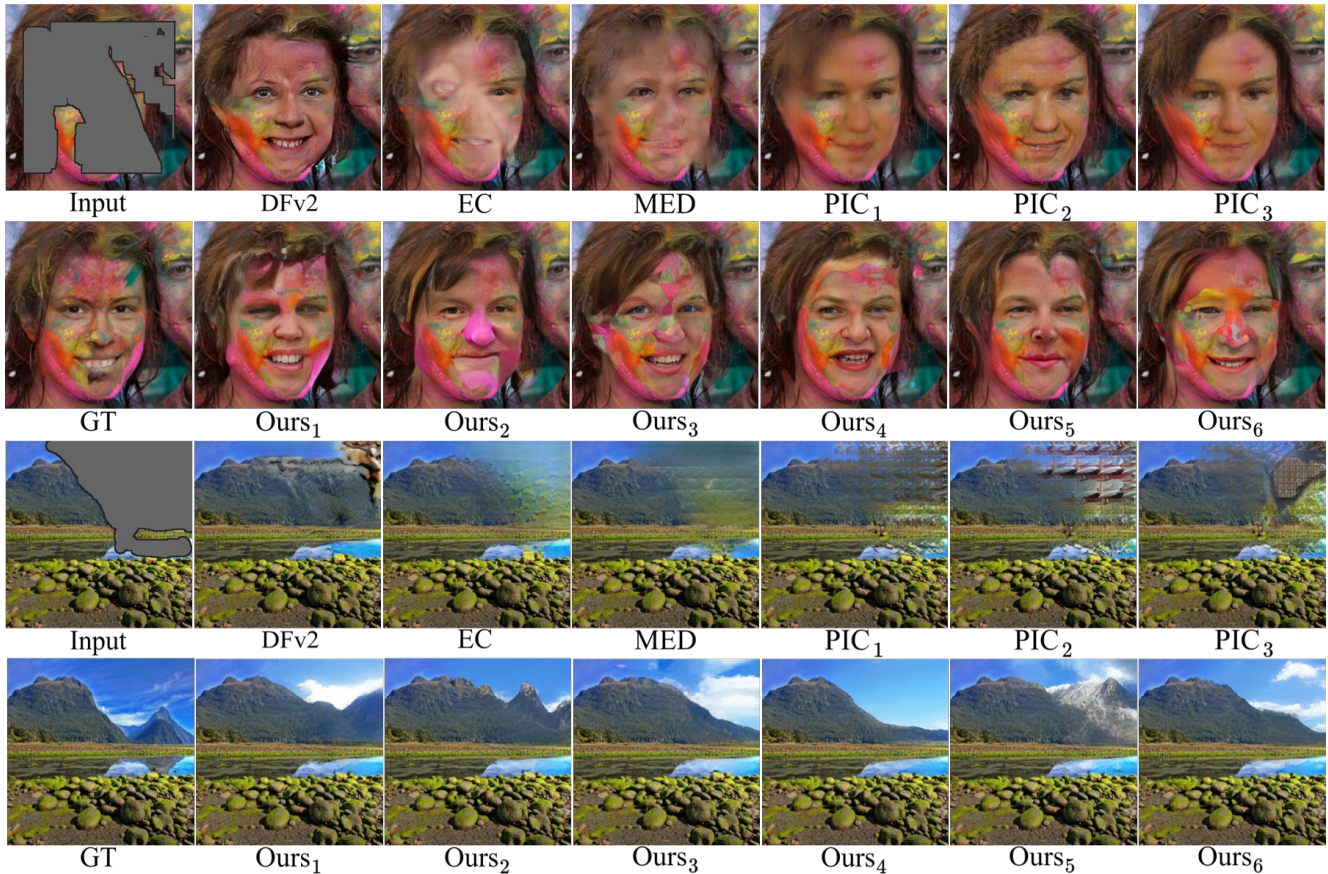
Figure 4: **Qualitative comparison with state-of-the-art methods on FFHQ, Places2 dataset.** The completion results of our method are with better quality and diversity.

The loss weights are set to $\alpha_1 = 1.0$ and $\alpha_2 = 0.1$ in all experiments. We also observe that involving instance normalization (IN) [31] will cause color inconsistency and severe artifacts during optimization. Therefore we remove all IN in the upsampling network.

4. Experiments

We present the details of implementation in Sec. 4.1, subsequently evaluate (Sec. 4.2) and delve into (Sec. 4.3) the proposed transformer-based image completion method. The pluralistic image completion experiments are conducted at 256×256 resolution on three datasets including *FFHQ* [16], *Places2* [42] and *ImageNet* [29]. We preserve 1K images from the whole *FFHQ* for testing, and use the original common training and test splits in rest datasets. The diversified irregular mask dataset provided by PConv [19] is employed for both training and evaluation.

4.1. Implementation Details

We control the scale of transformer architecture by balancing the representation ability and size of dataset concurrently. The discrete sequence length $\mathbb{L}$ is set to 48×48 for *FFHQ*. Limited by the computational resources, we de-

crease feasible $\mathbb{L}$ to 32×32 on large-scale datasets *Places2* and *ImageNet*. The detailed configurations of different transformer models are attached in supplementary material.

We use 8 Tesla V100 GPUs for *FFHQ* with batch size 64, 4×8 Tesla V100 GPUs for *Places2* and *ImageNet* with batch size 256 to train the transformers until convergence. We optimize the transformer using AdamW [24] with $\beta_1 = 0.9$, $\beta_2 = 0.95$. The learning rate is warmed up from 0 to 3e-4 linearly in the first epoch, then decays to 0 via cosine scheduler in rest iterations. No extra weight decay and dropout strategy are employed in the model. To train the guided upsampling network we use Adam [17] optimizer with fixed learning rate 1e-4, $\beta_1 = 0.0$ and $\beta_2 = 0.9$. During optimization, the weights of different loss terms are set to fixed value described in Sec. 3.2 empirically.

4.2. Results

Our method is compared against the following state-of-the-art inpainting algorithms: Edge-Connect (EC) [25], DeepFillv2 (DFv2) [37], MED [20] and PIC [41] using the official pre-trained models. We also fully train these models on *FFHQ* dataset for fair comparison.

Qualitative Comparisons We qualitatively compare the

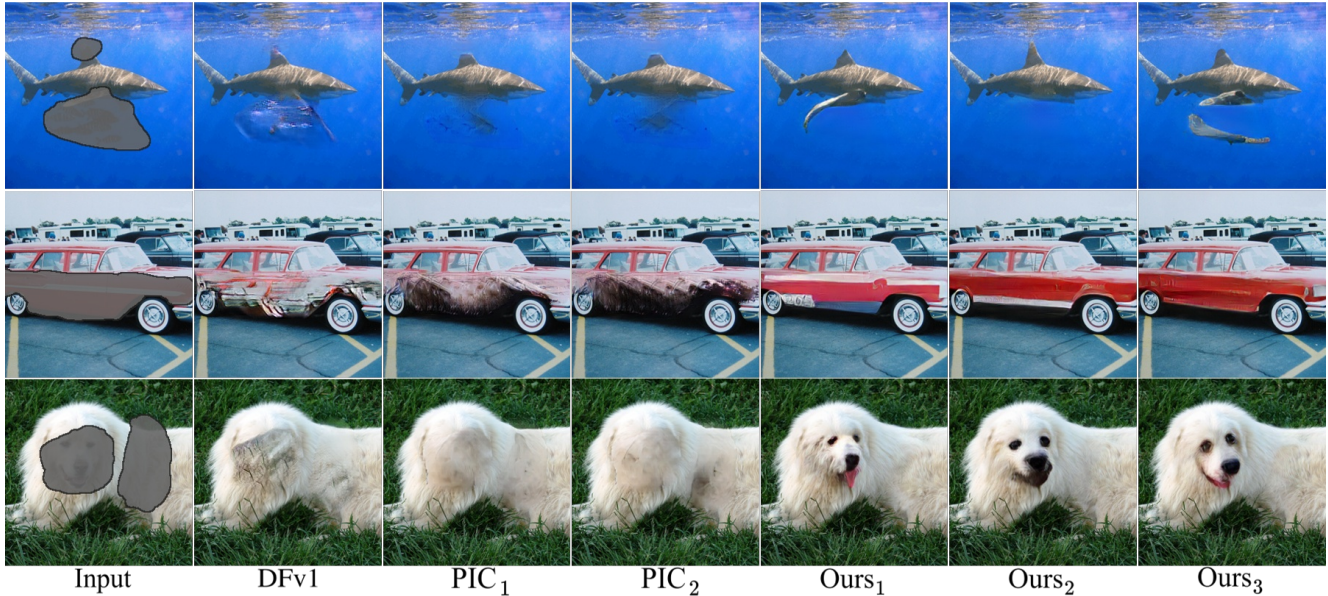| Input | DFv1 | PIC$_1$ | PIC$_2$ | Ours$_1$ | Ours$_2$ | Ours$_3$ |

Figure 5: **Qualitative comparison with state-of-the-art methods on ImageNet dataset.** More qualitative examples are shown in supplementary materials.

results with other baselines in this section. We adopt the sampling strategy introduced in Sec. 3.1 with $\mathcal{K}$=50 to generate 20 solutions in parallel, then select the top 6 results ranked by the discriminator score of the upsampling network following PIC [41]. All reported results of our method are direct outputs from the trained models without extra post-processing steps.

We show the results on the *FFHQ* and *Places2* datasets in Figure. 4. Specifically, EC [25] and MED [20] could generally produce the basic components of missing regions, but the absence of texture details makes their results non-photorealistic. DeepFillv2 [37], which is based on a multi-stage restoration framework, could generate sharp details. However, severe artifacts appear when mask regions are relatively large. Moreover, their method could only produce a single solution for each input. As the state-of-the-art diversified image inpainting method, PIC [41] tends to generate over-smooth contents and strange patterns, meanwhile the semantic-reasonable variation is limited to a small range. Compared to these methods, ours is superior in both photo-realism and diversity. We further show the comparison with DeepFillv1 (DFv1) [36] and PIC [41] on ImageNet dataset in Figure. 5. In this challenging setting, full CNN-based methods could not understand the global context well, resulting in unreasonable completion. When meeting large masks, they even could not maintain the accurate structure as shown in the second row of Figure. 5. In comparison, our method gives superior results, which demonstrates the exceptional generalization ability of our method on large-scale datasets.

Quantitative Comparisons We numerically compare our method with other baselines in Table. 1 and Table. 2. The peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and relative ℓ_1 (MAE) are used to compare the low-level differences between the recovered output and the ground truth, which is more suitable to measure the mask setting of small ratio. To evaluate the larger area missing, we adopt Fréchet Inception Distance (FID) [13], which calculates the feature distribution distance between completion results and natural images. Since our method could produce multiple solutions, we need to find one examplar to calculate the mentioned metrics. Unlike PIC [41], which selects the results with high ranking discriminator scores for each sample, here we directly provide stochastic sampling results while $\mathcal{K}$=50 to demonstrate its generalization capability. Furthermore, we also provide deterministic sampling results given $\mathcal{K}$=1 in Table. 1. It can be seen that our method with top-1 sampling achieves superior results compared with other competitors in almost all metrics. And in the case of relatively large mask regions, top-50 sampling leads to slightly better FID scores. On the ImageNet dataset, as shown in Table. 2, our method outperforms PIC by a considerable margin, especially in FID metrics (more than 41.2 for large masks).

User Study To better evaluate the subjective quality, we further conduct a user study to compare our method with other baselines. Specifically, we randomly select 30 masked images from the test set. For a test image, we use each method to generate one completion result and ask the participant to rank the total five results from the highest photo-realism to the lowest photorealism. We have collected answers from 28 participants and calculate the ratios of each method being selected as top 1,2,3, with the statistics shown in Figure. 6. Our method is 73.70% more likely to be picked

Dataset		FFHQ [16]				Places2 [42]			
Method	Mask Ratio	PSNR↑	SSIM↑	MAE↓	FID↓	PSNR↑	SSIM↑	MAE↓	FID↓
DFv2 [37]	20%-40%	25.868	0.922	0.0231	16.278	26.533	0.881	0.0215	24.763
EC [25]		26.901	0.938	0.0209	14.276	26.520	0.880	0.0220	25.642
PIC [41]		26.781	0.933	0.0215	14.513	26.099	0.865	0.0236	26.393
MED [20]		26.325	0.922	0.0230	14.791	26.469	0.877	0.0224	26.977
Ours		27.922	0.948	0.0208	10.995	26.503	0.880	0.0244	21.598
Ours†		**28.242**	**0.952**	**0.0155**	**10.515**	**26.712**	**0.884**	**0.0198**	**20.431**
DFv2 [37]	40%-60%	21.108	0.802	0.0510	28.711	22.192	0.729	0.0440	39.017
EC [25]		21.368	0.780	0.0510	30.499	22.225	0.731	0.0438	39.271
PIC [41]		21.723	0.811	0.0488	25.031	21.498	0.680	0.0507	49.093
MED [20]		20.765	0.763	0.0592	34.148	22.271	0.717	0.0457	45.455
Ours		22.613	0.845	0.0445	**20.024**	22.215	0.724	0.0431	**33.853**
Ours†		**23.076**	**0.864**	**0.0371**	20.843	**22.635**	**0.739**	**0.0401**	34.206
DFv2 [37]	Random	24.962	0.882	0.0310	19.506	25.692	0.834	0.0280	29.981
EC [25]		25.908	0.882	0.0301	17.039	25.510	0.831	0.0293	30.130
PIC [41]		25.580	0.889	0.0303	17.364	25.035	0.806	0.0315	33.472
MED [20]		25.118	0.867	0.0349	19.644	25.632	0.827	0.0291	31.395
Ours		26.681	0.910	0.0292	14.529	25.788	0.832	0.0267	**25.420**
Ours†		**27.157**	**0.922**	**0.0223**	**14.039**	**25.982**	**0.839**	**0.0254**	25.985

Table 1: **Quantitative results on FFHQ and Places2 datasets with different mask ratios.** Ours: Default top-50 sampling. Ours†: Top-1 sampling.

Method	Mask Ratio	PSNR↑	SSIM↑	MAE↓	FID↓
PIC [41]	20%-40%	24.010	0.867	0.0319	47.750
Ours		**24.757**	**0.888**	**0.0263**	**28.818**
PIC [41]	40%-60%	18.843	0.642	0.0756	101.278
Ours		**20.135**	**0.721**	**0.0585**	**59.486**
PIC [41]	Random	22.711	0.791	0.0462	59.428
Ours		**23.775**	**0.835**	**0.0358**	**35.842**

Table 2: **Quantitative comparison with PIC on ImageNet dataset.**

Figure 6: **Results of user study.**

as the first rank, which demonstrates its clear advantage.

4.3. Analysis

Diversity. We calculate the average LPIPS distance [38] between pairs of randomly-sampled outputs from the same input to measure the completion diversity following Zhu *et al.* [43]. Specifically, we leverage 1K input images and sample 5 output pairs per input in different mask ratios. And the LPIPS is computed based on the deep features of VGG [30] model pre-trained on ImageNet. The diversity scores are shown in Figure. 7. Since the completion with diverse but
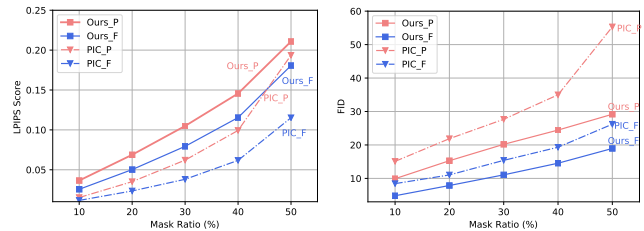

Figure 7: **Left:** Diversity curve. **Right:** FID curve. **P** and **F** denote *Places2* and *FFHQ* dataset respectively.

meaningless contents will also lead to high LPIPS scores, we simultaneously provide the FID score of each level computed between the whole sampled results (10K) and natural images in the right part of Figure. 7. Our method achieves better diversity in all cases. Furthermore, in the max mask ratio of Places2, although PIC [41] approximates the diversity of our method, the perceptual quality of our completion outperforms PIC [41] by a large margin.

Robustness for the completion of extremely large holes. To further understand the ability of the transformer, we conduct extra experiments on the setting of extremely large holes, which means only very limited pixels are visible. Although both the transformer and upsampling network are only trained using a dataset of PConv [19] (max mask ratio 60%), our method generalizes fairly well to this difficult setting. In Figure. 8, almost all the baselines fail with large missing regions, while our method could produce high-quality and diversified completion results.

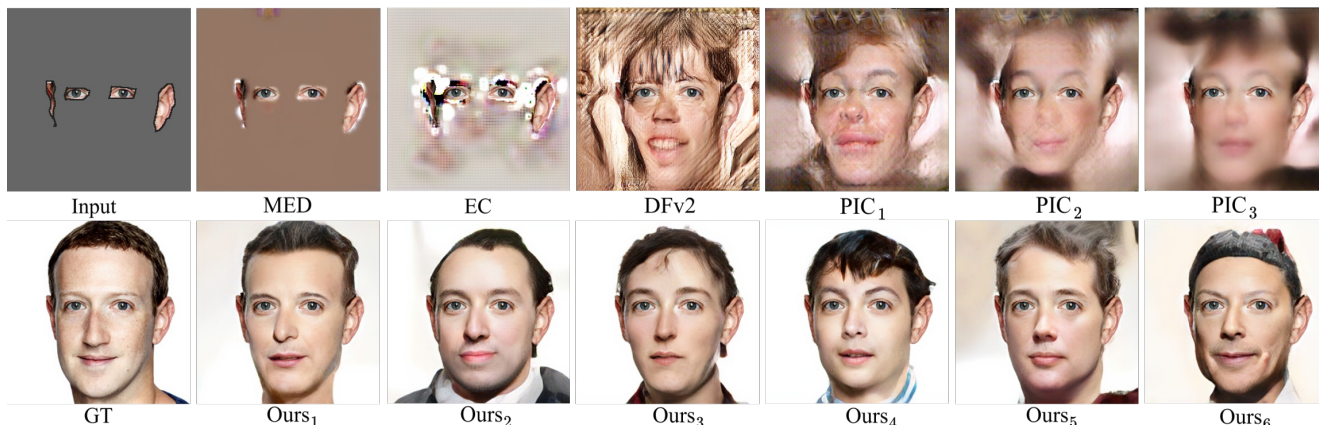**If the transformer could better understand global struc-**

Figure 8: **Image completion results of large-scale masks.** It could be noted that all the compared baselines struggle to generate plausible contents.
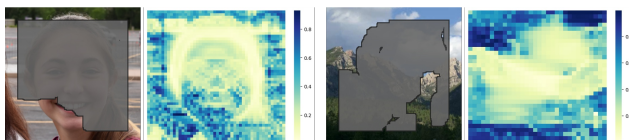

Figure 9: **Visualization of probability map generated by transformer.** Higher confidence denotes lower uncertainty.

Figure 10: **Completion of basic geometric shape.** All compared models are trained on ImageNet. Ours$_{ap}$: Appearance prior reconstructed from transformer.

ture than CNN? To answer this question, we conduct the completion experiments on some geometric primitives. Specifically, we ask our transformer-based method and other full CNN methods, *i.e.* DeepFillv1 [36] and PIC [41], trained on ImageNet to recover the missing parts of pentagram shape in Figure. 10. As expected, all full CNN methods fail to reconstruct the missing shape, which may be caused by the locality of the convolution kernel. In contrast, the transformer could easily reconstruct the right geometry in low-dimensional discrete space. Based on such accurate appearance prior, the upsampling network could more effectively render the original resolution results finally.

Visualization of probability map. Intuitively, since the contour of missing regions is contiguous to existing pixels, the completion confidence should gradually decrease from the mask boundary to the interior region. The lower confidence corresponds to more diversified possible contents. To verify this point, we plot the probability map in Figure. 9 , where each pixel denotes the maximum probability of visual vocabulary generated by the transformer. And we have some interesting observations: 1) In the right part of Figure. 9, the uncertainty is indeed increasing from outside to inside. 2) For the portrait completion example, the uncertainty of face regions is overall lower than hair parts. The underlying reason is the seen parts of the face constrain the diversity of other regions to some degree. 3) The probability of the right cheek of the portrait example is highest among the rest mask regions, which indicates that the transformer captures the symmetric property.
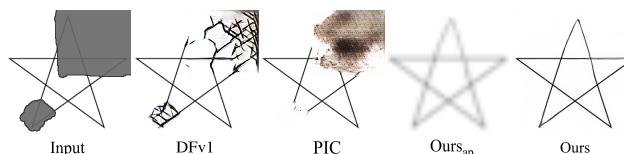
5. Concluding Remarks

There is a long-existing dilemma in the image completion area to achieve both enough diversity and photorealistic quality. Existing attempts mostly optimize the variational lower-bound through a full CNN architecture, which not only limits the generation quality but is also difficult to render natural variations. In this paper, we first propose to bring the best of both transformers and CNNs to achieve high-fidelity free-form pluralistic image completion. Extensive experiments are conducted to demonstrate that the superiority of our method compared with state-of-the-art fully convolutional approaches, including large performance gain on regular evaluations setting, more diversified and vivid results, and exceptional generalization ability.

We have shown the the great potentials of bi-directional transformers in image completion areas. While these results are encouraging, we believe there are still some challenges could be considered in the future, including inference efficiency, sampling strategy or architecture design. We recommend the audience to check the provided supplementary material[1] for more intensive analysis.

[1] http://raywzy.com/ICT/supple.pdf

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: a randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (TOG)*, 28(3):1–11, 2009.

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

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020.

[5] Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. 2020.

[6] Donghyeon Cho, Jinsun Park, Tae-Hyun Oh, Yu-Wing Tai, and In So Kweon. Weakly-and self-supervised learning for content-aware deep image retargeting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4558–4567, 2017.

[7] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: Combining inconsistent images using patch-based synthesis. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Alexei A Efros and William T Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346, 2001.

[12] James Hays and Alexei A Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (TOG)*, 26(3):4–es, 2007.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[14] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. 2017.

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

[16] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

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

[18] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. 2017.

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

[20] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. *arXiv preprint arXiv:2007.06929*, 2020.

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

[22] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9371–9381, 2021.

[23] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, Jing Liao, Bin Jiang, and Wei Liu. Deflocnet: Deep image editing via flexible low-level controls. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10765–10774, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019.

[26] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International Conference on Machine Learning*, pages 4055–4064, 2018.

[27] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. 2016.

[28] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[32] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29:4790–4798, 2016.

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

[34] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2747–2757, 2020.

[35] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Old photo restoration via deep latent space translation. *arXiv preprint arXiv:2009.07047*, 2020.

[36] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. 2018.

[37] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

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

[39] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5741–5750, 2020.

[40] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, 2017.

[41] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019.

[42] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[43] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *arXiv preprint arXiv:1711.11586*, 2017.