Bridging Global Context Interactions for High-Fidelity Image Completion

Chuanxia Zheng[1] Tat-Jen Cham[2] Jianfei Cai[1] Dinh Phung[1]
[1]Department of Data Science & AI, Monash University, Australia
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
chuanxia.zheng@monash.edu, ASTJCham@ntu.edu.sg, {Jianfei.Cai,dinh.phung}@monash.edu

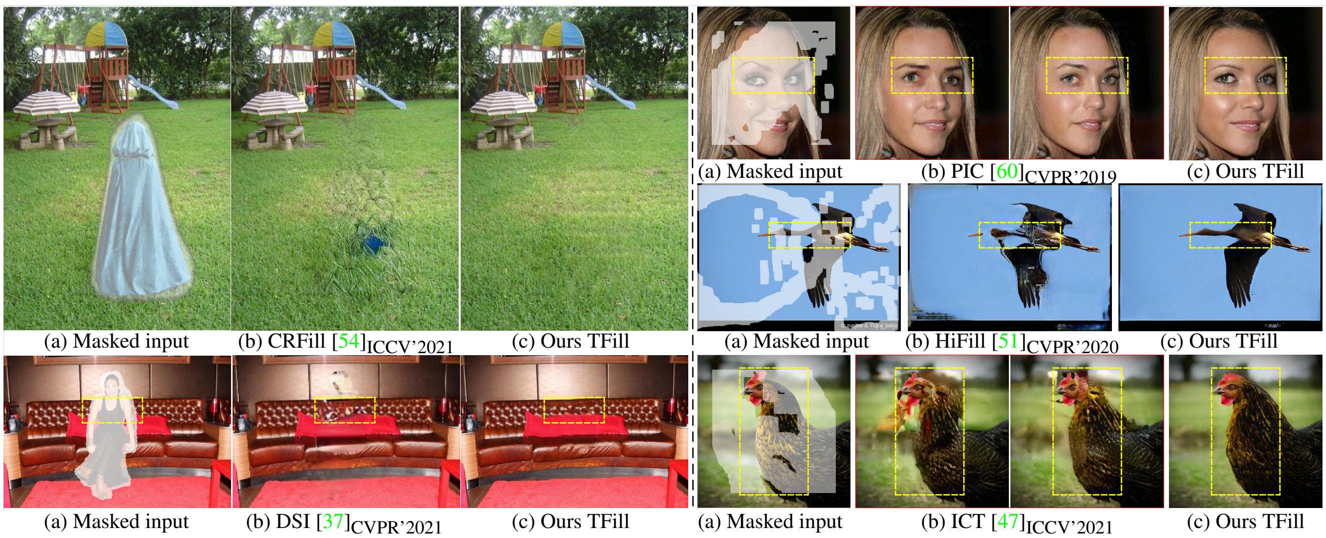


Figure 1. **Example completion results of our method on different sceneries with various masks** (missing regions shown in white, a transparency ratio is set for better visualization). Our TFill model not only effectively removes large objects (left), but also infers reasonable contents and plausible appearances for semantical image completion on various settings (right). (Zoom in to see the details.)

Abstract

Bridging global context interactions correctly is important for high-fidelity image completion with large masks. Previous methods attempting this via deep or large receptive field (RF) convolutions cannot escape from the dominance of nearby interactions, which may be inferior. In this paper, we propose to treat image completion as a directionless sequence-to-sequence prediction task, and deploy a transformer to directly capture long-range dependence. Crucially, we employ a restrictive CNN with small and non-overlapping RF for weighted token representation, which allows the transformer to explicitly model the long-range visible context relations with equal importance in all layers, without implicitly confounding neighboring tokens when larger RFs are used. To improve appearance consistency between visible and generated regions, a novel attention-aware layer (AAL) is introduced to better exploit distantly related high-frequency features. Overall, extensive experiments demonstrate superior performance compared to state-of-the-art methods on several datasets. Code is available at https://github.com/lyndonzheng/TFill.

1. Introduction

Image completion refers to the task of filling reasonable content with photorealistic appearance into missing regions, conditioned on partially visible information (as shown in Fig. 1). Earlier methods infer the pixels of missing regions by propagating pieces from neighboring visible regions [1–3, 9], while more recent ones directly learn to generate content and appearance using deep neural networks [17, 28, 29, 35–37, 47, 51, 52, 54, 60].

A main challenge in this task is the requirement of *bridging and exploiting visible information globally, after it had been degraded by arbitrary masks*. As depicted on the left of Fig. 1, when the entire person is masked, the natural expectation is to complete the masked area based on the visible background context. In contrast, on the right of Fig. 1, when large free-form irregular masks cover the main parts but leave the partial information visible, it is necessary but highly challenging to correctly capture *long-range* dependencies between the separated foreground regions, so that the masked area can be completed in not just a photorealistic, but also semantically correct, manner.

To achieve this goal, several *two-stage* approaches [35, 37, 51, 52, 54] have been proposed, consisting of a *content inference network* and an *appearance refinement network*. They typically infer a coarse image/edge/semantic map based on globally visible context in a first phase, and then fill in visually realistic appearances in a second phase. However, this global perception is achieved by repeating *local* convolutional operations, which have several limitations. First, due to the translation equivariance, the information flow tends to be predominantly local, with global information only shared gradually through *heat-like propagation* across multiple layers. Second, during inference, the elements between adjacent layers are connected via learned but *fixed* weights, rather than input-dependent adaptive weightings. These issues mean long-distance messages are only delivered inefficiently in a very deep layer, resulting in a strong inclination for the network to fill holes based on nearby rather than distant visible pixels (*cf*. Fig. 1).

In this paper, we propose an alternative perspective by treating image completion as a *directionless sequence-to-sequence* prediction task. In particular, instead of modeling the global context using deeply stacked convolutional layers, we design a new content inference model, called TFill, that uses a **T**ransformer-based architecture to **Fill** reasonable content into the missing holes. An important insight here is that a transformer directly exploits long-range dependencies at every encoder layer through the attention mechanism, which *creates an equal flowing opportunity for all visible pixels, regardless of their relative spatial positions* (Fig. 4 (c)). This reduces the proximity-dominant influence that can lead to semantically incoherent results.

However, it remains a challenge to directly apply these transformer models to visual generation tasks. In particular, unlike in NLP where each word is naturally treated as a vector for token embedding [10, 39, 40, 46], it is unclear *what a good token representation should be for a visual task*. If we use every pixel as a token, the memory cost will make this infeasible except for very small *downsampled* images [8, 47]. To mitigate this issue, our model embeds the masked image into an intermediate latent space for token representation, an approach also broadly taken by recent vision transformers [6, 12, 49, 62, 64]. However, unlike these models that use conventional CNN-based encoders to embed the tokens, *without considering the visible information flow in image completion*, we propose a *restrictive CNN* for token representation, which has a profound influence on how the visible information is connected in the network. To do so, we ensure the individual tokens represent visible information independently, each within a *small* and *non-overlapping* patch, and forces *the long-range context relationships between tokens to be explicitly and co-equally perceived in every transformer encoder layer*. As a result, each masked pixel will *not* be gradually affected by neighboring visible pixels.

While the proposed transformer-based architecture can achieve better results than state-of-the-art methods [12, 51, 52, 60], by itself it only works for a *fixed* sequence length because of the position embedding (Fig. 2(a)). To allow our approach to flexibly scale to images of arbitrary sizes, *especially at high resolution*, a fully convolutional network (Fig. 2(b)) is subsequently applied to refine the visual appearance, building upon the coarse content previously inferred. A novel **A**ttention-**A**ware **L**ayer (AAL) is inserted between the encoder and decoder that adaptively balances the attention paid to visible and generated content, leading to semantically superior feature transfer (Figs. 5 and 9).

We highlight our main contributions as follows: **1)** A *restrictive* CNN head is introduced for individual *weighted* token representation, which mitigates the proximity influence when propagating visible information to missing holes. **2)** Through a transformer-based architecture, the long-range interactions between these tokens are explicitly modeled, in which the masked tokens are perceptive of other visible tokens with equal opportunity, regardless of their positions. **3)** A novel attention-aware layer with adaptive attention balancing is introduced in a refined stage to obtain higher quality and resolution results. **4)** Finally, extensive experiments demonstrate that the proposed model outperforms the existing state-of-the-art image completion models.

2. Related Work

Image Completion: Traditional image completion (also known as "image inpainting" [3]) methods, like diffusion-based [1, 4, 26] and patch-based [2, 9, 18], mainly focus on background completion, by directly copying and propagating the background pixels to masked regions.

Driven by the advances of GANs [13], CGANs [34] and VAEs [25], a series of CNN-based methods [17, 28, 35, 36, 44, 52, 59, 60] have been proposed to hallucinate semantic meaningful content. In particular, Pathak *et al*. [36] introduced GANs into image completion for large holes. Iizuka *et al*. [17] extended [36] to random regular mask. Yu *et al*. [52] combined the patch-based idea into learning-based architecture, which is followed by [42, 43, 51, 54, 55, 60]. Liu *et al*. [28] addressed random irregular masks. Zheng *et al*. [60, 61] introduced a pluralistic image completion task, aiming to generate multiple and diverse results, which is followed by [30, 37, 47, 58]. Nazeri *et al*. [35] brought the auxiliary edge information for image completion. Then, more auxiliary information were combined into this task, *e.g*. Faceshape [38], DeepFill v2 [53], SC-FEGAN [20], SWAP [27], and MST [5]. Most of these models are built upon on a CNN-based architecture, in which the masked re-

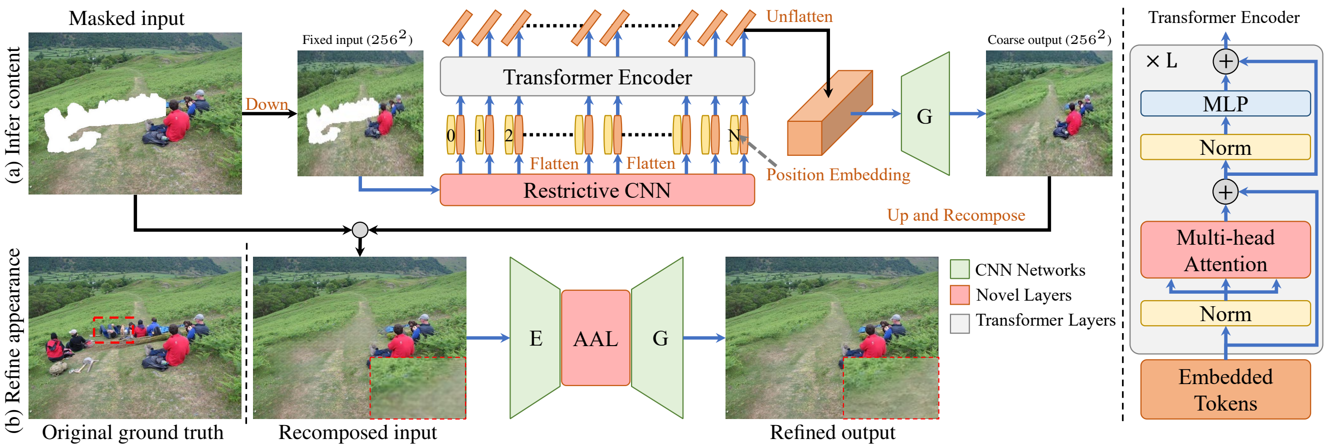


Figure 2. **The overall pipeline of our proposed method**. (a) Masked input is resized to a fixed resolution (256×256) and it is then fed into a transformer encoder to generate semantically correct content. (b) The inferred content is merged with the original high-resolution image and passed to a refinement network with an **A**ttention-**A**ware **L**ayer (**AAL**) to transfer high-quality information from both visible and generated regions. Note the recomposed input has repeating artifacts, which are ameliorated in our refined network.

gions are gradually affected by the neighboring visible pixels. Our model will solve this problem by utilizing a transformer to directly model the global context dependencies.

Visual Transformer: The Transformer was firstly proposed by Vaswani *et al*. [46] for machine translation. Inspired by the dramatic success of transformers in NLP [10, 40], recent works have explored applying a standard transformer for vision tasks [32], such as image classification [8, 11, 14], object detection [6, 64], semantic segmentation [48, 62], image generation and translation [8, 12, 16, 19], and completion [31, 47]. Many of these embed tokens using methods shown in Fig. 3(a)-(c), without considering the specific information flow in image completion. In contrast, our *restrictive CNN* is particularly well suited due to its compact representation in the form of local patches.

Context Attention: Context attention [52] is a specific cross-attention that aims to copy high-frequency details from high-resolution encoded features to generate high-quality images. It has recently been widely applied in image completion [42, 50–52, 54, 60]. However, the existing works mainly copy from visible regions [42, 50–52, 54], which is not possible for newly generated content. In addition, our AAL automatically selects features from both "visible" encoded and "missing" generated features, instead of selecting through *fixed* weights [60].

3. Methods

Given a masked image $\mathbf{I}_m$, degraded from a real image $\mathbf{I}$ by masks, our goal is to learn a model Φ to infer semantically reasonable content for missing regions, as well as filling in with visually realistic appearance.

To achieve this, our framework, illustrated in Fig. 2, consists of a content inference network (TFill-*Coarse*, Fig. 2(a)) and an appearance refinement network (TFill-*Refined*, Fig. 2(b)). The former is responsible for capturing the global context through a transformer encoder. The embedded tokens have small receptive fields (RF) and limited capacity, *preventing masked pixels' states from being implicitly dominated by visible pixels nearby than far*. While similar transformer-based architectures have recently been explored for visual tasks [6–8, 11, 12, 47, 48, 62, 64], we discover *how the token representation has a profound effect on the flow of visible information in image completion, in spite of the supposedly global reach of transformers*. The latter network is designed to refine appearance by utilizing high-resolution visible features globally, and also frees the limitation to fixed sizes.

3.1. Content Inference Network: TFill-*Coarse*

Our TFill-*Coarse* depends on the `self-attention` module in a transformer-encoder to *equally* perceiving global visible context for the completed content generation. Considering the *fixed* length position embedding and dramatically increased computational cost, we first downsample images with arbitrary sizes to a *fixed* size, *e.g*. 256×256. However, it is still *not* feasible to run a transformer model if we directly *flatten* image pixels into a 2D sequence.

To obtain a practicable number of visual tokens, different embedding methods (Fig. 3(a)-(c)) have been used in current visual transformer-based works [6, 8, 11, 12, 16, 19, 47, 48, 62, 64]. These visual tokens' RF is either as small as a pixel (*e.g*. iGPT [8]) that loses important context details due to the large-scale downsampling, or is as large as the full image size (*e.g*. VQGAN [12]) that has firstly been gradually influenced by neighboring pixels in deep CNN

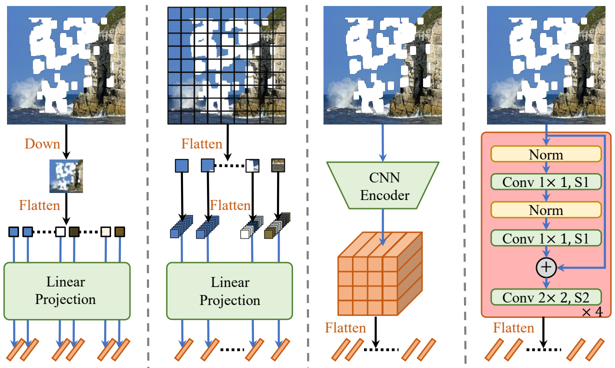


Figure 3. **Token representation.** (a) Pixel to token. (b) Patch to token. (c) Feature to token. (d) Restrictive **R**eceptive **F**ield (RF) feature to token. Note our token has a small and non-overlapping RF like VIT [11], but uses a stacked (x4) CNN embedding. Each token represents locally isolated patches, leaving the global context relationship to be cleanly modeled in a transformer encoder.

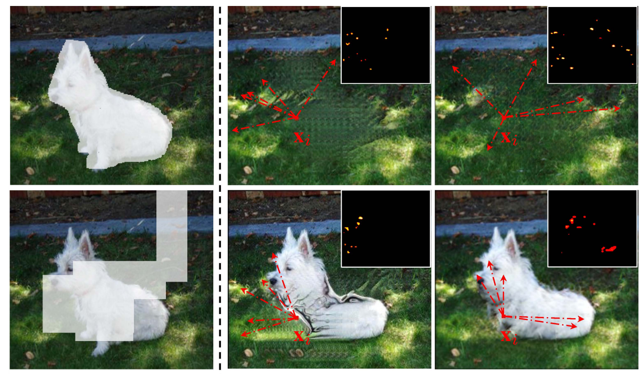


Figure 4. **An example of information flow in image completion.** The position $\mathbf{x}_i$'s response (flow) is calculated by inferring the *Jacobian* matrix between it to all pixels in the given masked input. Here, only the highest flows are shown. Our TFill correctly captures long-range visible context flow, even with a large mask splitting two semantically important zones.

layers. While patch embedding [11] achieves impressive performance in many tasks, one-layer linear projection is still not good enough [49].

Restrictive CNN: In contrast to these methods, our token representation is extracted using a *restrictive CNN* (Fig. 3(d)) in 4 blocks. In each block, the 1×1 filter and `layernorm` is applied for non-linear projection, followed by a partial convolution layer [28] that uses a $2{\times}2$ filter with stride 2 to extract visible information. In particular, if half of the regions in a window are masked, we only embed the other 50% comprising visible pixels as our token representation, and establish an initial weight of 0.5 for the next *weighted* self-attention layer. To do this, we ensure each token represents only the visible information in a local patch, *leaving the long-range dependencies to be explicitly modeled by a transformer*, without cross-contamination from implicit correlation due to larger CNN RF.

In fact, some latest works also begin to explore the influence of different token embeddings. Swin [32] used shift windows to get multi-scales embedded features. ViT$_c$ [49] demonstrated an early CNN token embedding is important for visual transformer. However, they do not consider information flowing from visible to masked regions. When a large RF is applied into a deep CNN embedding, the masked holes will be gradually determined by the neighboring visible pixels. In Fig. 4, we empirically show this is precisely the case for prior CNN-based models. Because masked regions originally hold zero values, they will take the neighboring visible pixels as a filled and reasonable value for the next layer. In contrast, as the small patch is directly embedded using local visible information with important *weight*, the proposed *restrictive* CNN is better suited for image completion task.

***Weighted* Self-Attention Layer:** To further *bias* the important visible values, we replace the self-attention layer with a *weighted* self-attention layer, in which a weight is applied to scale the attention scores. The initial weight $w^{(1)} \in (0.02, 1.0]$ is obtained by calculating the fraction of visible pixels in a small patch, *e.g.* $192/(16 \times 16)$ means 192 pixels in the 16×16 patch are visible. It will then be gradually amplified by updating $w^{(i+1)} \leftarrow \sqrt{w^{(i)}}$ after every encoder layer, to *reflect* visible information flow. This initial ratio for each token is efficiently implemented in our restrictive CNN encoder.

CNN-based Decoder: Following existing works [51, 52, 60], a gradual upsampling decoder is implemented to generate photorealistic images. Instead of sequentially generating tokens, our model directly predict all tokens in one step, resulting in a much faster testing time than existing transformer-based generation networks [8,12,47] (Table 3).

3.2. Appearance Refinement Network: TFill-*Refined*

Although the proposed TFill-*Coarse* model correctly infers superiorly reasonable content (shown in Figs. A.1, A.2, and A.3) by equally utilizing the global visible context in every layer, two limitations remain. First, it is *not* suitable for high-resolution input due to the *fixed* length position embedding. Second, the realistic completed results may *not be fully consistent with the original visible appearances*, *e.g.* the completed eye in Fig. 5 (c).

Attention-Aware Layer (AAL): To mitigate these issues, a refinement network, trained on high-resolution images, is proposed (Fig. 2 (b)). In particular, to further utilize the visible high-frequency details in global, an **A**ttention-

(a) Ground truth (b) Masked input

(c) TFill-*Coarse* (d) TFill-*Refined*

Figure 5. **Coarse and Refined results**. (a) Ground truth. (b) Masked input. (c) Coarse output. (d) Refined output. The refinement network not only increases image quality to a high resolution (256^2 *vs* 512^2), but also encourages the left eyeball to be consistent with the visible right eyeball using our attention-aware layer.

Aware **L**ayer (AAL) is designed to *copy long-range information from both encoded and decoded features*.

As depicted in Fig. 6, given a decoded feature $\mathbf{x}_d$, we first calculate the attention score of:

$$\mathbf{A} = \phi(\mathbf{x}_d)^\intercal \theta(\mathbf{x}_d), \tag{1}$$

where $\mathbf{A}_{ij}$ represents the similarity of the i^{th} feature to the j^{th} feature, and ϕ, θ are 1×1 convolution filters.

Interestingly, we discover that using $\mathbf{A}$ directly in a standard self-attention layer is suboptimal, because the $\mathbf{x}_d$ features for visible regions are generally distinct from those generated for masked regions. Consequently, *the attention tends to be insular*, with masked regions preferentially attending to masked regions, and vice versa. To avoid this problem, we explicitly handled the attention to visible regions separately from masked regions. So before `softmax` normalization, $\mathbf{A}$ is split into two parts: $\mathbf{A}_v$ — similarity to *visible* regions, and $\mathbf{A}_m$ — similarity to generated *masked* regions. Next, we get long-range dependencies via:

$$\mathbf{z}_v = \texttt{softmax}(\mathbf{A}_v)\mathbf{x}_e \quad , \quad \mathbf{z}_m = \texttt{softmax}(\mathbf{A}_m)\mathbf{x}_d \tag{2}$$

where $\mathbf{z}_v$ contains features of contextual flow [52] for copying high-frequency details from the encoded high-resolution features $\mathbf{x}_e$ to masked regions, while $\mathbf{z}_m$ has features from the self-attention that is used in SAGAN [56] for high-quality image generation.

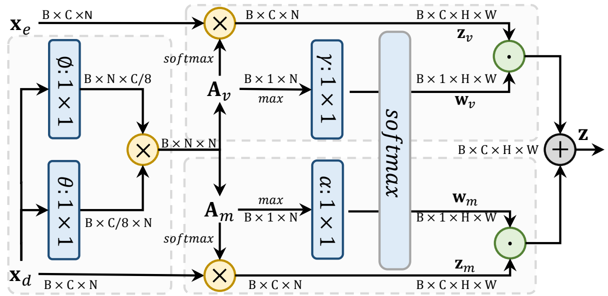


Figure 6. **Attention-aware layer.** The feature maps are shown as tensors. "$\bigotimes$" denotes matrix multiplication, "$\bigodot$" denotes element-wise multiplication and "$\bigoplus$" is element-wise sum. The blue boxes denote 1×1 convolution filters that are learned.

Instead of learning *fixed* weights [60] to combine $\mathbf{z}_v$ and $\mathbf{z}_m$, we learn the *weights mapping* based on the largest attention score in each position. Specifically, we first obtain the largest attention score of $\mathbf{A}_v$ and $\mathbf{A}_m$, respectively. Then, we use the 1×1 filter γ and α to *modulate* the ratio of the weights. `Softmax` normalization is applied to ensure $\mathbf{w}_v + \mathbf{w}_m = 1$ in every spatial position:

$$[\mathbf{w}_v, \mathbf{w}_w] = \texttt{softmax}([\gamma(\texttt{max}(\mathbf{A}_v)), \alpha(\texttt{max}(\mathbf{A}_m)]]) \tag{3}$$

where `max` is executed on the attention score channel. Finally, an attention-balanced output $\mathbf{z}$ is obtained by:

$$\mathbf{z} = \mathbf{w}_v \cdot \mathbf{z}_v + \mathbf{w}_m \cdot \mathbf{z}_m \tag{4}$$

where $\mathbf{w}_v, \mathbf{w}_m \in \mathbb{R}^{B\times1\times H\times W}$ hold different values for various positions, dependent on the largest attention scores in the visible and masked regions, respectively.

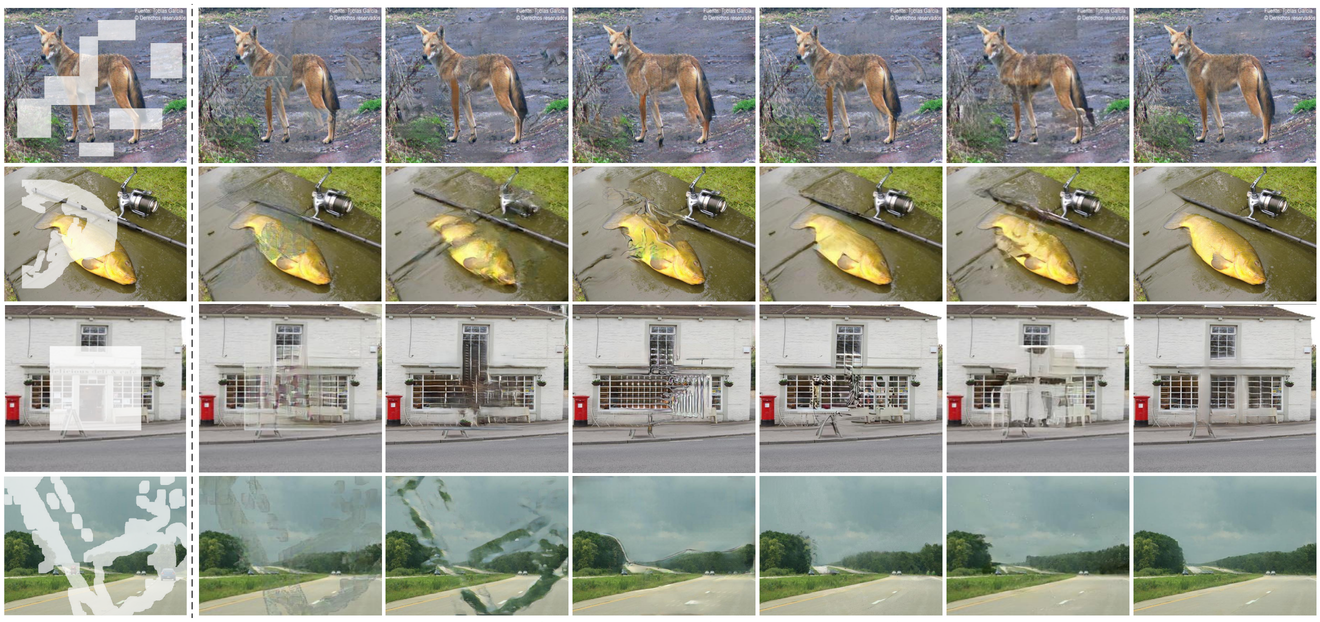
4. Experiments

4.1. Experimental Details

Datasets: We evaluated the proposed TFill model with arbitrary mask types on various datasets, including CelebA-HQ [22, 33], FFHQ [23], Places2 [63], and ImageNet [41].

Metrics: Following existing works [35, 47, 61], we mainly reported the traditional patch-level image quality metrics, including peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM), and the latest learned feature-level LPIPS [57] and FID [15] metrics.

Implementation Details: Our model is trained in two stages: **1)** the TFill-*Coarse* is first trained for 256×256 resolution; and **2)** the TFill-*Refined* is then trained for 512×512 resolution. Unless other noted, TFill indicates the whole model in the paper. Both networks are optimized using the loss $L = L_{pixel} + L_{per} + L_{GAN}$, where L_{pixel} is the ℓ_1 reconstruction loss, L_{per} is the perceptual loss [21], and L_{GAN} is the discriminator loss [13]. More implementation details are provided in Appendix C.



(a) Masked input (b) GL [17] (c) DeepFillv2 [53] (d) HiFill [51] (e) CRFill [54] (f) ICT [47] (g) Ours TFill

Figure 7. **Qualitative comparison on various datasets with free-form masks.** Here, we show results for ImageNet [41] (top two examples) and Places2 [63] (bottom two examples). Our model generated more reasonable object and scene structures, with better visual results. Please zoom in to see the details. More comparisons are provided in Figs. A.4, A.5, and A.6.

	PSNR↑			SSIM↑			LPIPS↓			FID↓		
Mask Ratio	20-30%	30-40%	40-50%	20-30%	30-40%	40-50%	20-30%	30-40%	40-50%	20-30%	30-40%	40-50%
GL [17]$_{\text{SIGGRAPH'2017}}$	21.33	19.11	17.56	0.7672	0.6823	0.5987	0.1847	0.2535	0.3189	39.22	53.24	68.46
PIC [60]$_{\text{CVPR'2019}}$	24.44	22.32	20.71	0.8520	0.7850	0.7119	0.1183	0.1666	0.2245	21.62	29.59	41.60
DeepFillv2 [53]$_{\text{ICCV'2019}}$	23.58	21.50	19.94	0.8319	0.7712	0.7074	0.1234	0.1639	0.2079	23.18	28.87	35.21
HiFill [51]$_{\text{CVPR'2020}}$	22.54	20.15	18.48	0.7838	0.7057	0.6193	0.1632	0.2258	0.3053	26.89	38.40	56.24
CRFill [54]$_{\text{ICCV'2021}}$	24.38	21.95	20.44	0.8476	0.7983	0.7217	0.1189	0.1597	0.1993	17.58	23.05	29.97
ICT [47]$_{\text{ICCV'2021}}$	24.53	22.84	21.11	0.8599	0.7995	0.7228	0.1045	0.1563	0.1974	17.13	22.39	28.18
Ours TFill	**25.10**	**22.89**	**21.22**	**0.8686**	**0.8063**	**0.7391**	**0.0918**	**0.1328**	**0.1796**	**15.28**	**19.99**	**25.88**

Table 1. Quantitative comparisons on Places2 [63] with free-form masks [28]. Without bells and whistles, TFill outperformed all existing learning-based models. The results are reported on 256×256 resolution, as earlier works were trained only on this scale.

4.2. Main Results

We firstly compared with the following state-of-the-art image completion methods: GL [17]$_{\text{SIGGRAPH'2017}}$, DeepFillv2 [53]$_{\text{ICCV'2019}}$, PIC [60]$_{\text{CVPR'2019}}$, HiFill [51]$_{\text{CVPR'2020}}$, CRFill [54]$_{\text{ICCV'2021}}$, and ICT [47]$_{\text{ICCV'2021}}$ using their publicly released codes and models.

Quantitative Results: Table 1 shows quantitative evaluation results on Places2 [63], in which the images were degraded by free-form masks provided in the PConv [28] testing set. The mask ratio denotes the range of masking proportion applied to the images. The original mask ratios hold six levels, from 0 to 60%, increasing 10% for each level. Here, following ICT [47], we only compare the results on middle-level mask ratios. As can be seen, the proposed TFill model outperformed the CNN-based state-of-the-art models in all mask scales. Specifically, it achieves averaging relative 18.8% and 13.3% improvements for LPIPS and FID scores, respectively. While the latest ICT [47] utilized the transformer architecture with much more blocks and more expensive computer cost, they downsampled the original image into 32×32, or 48×48 resolution, and then embedded each pixel as a token, resulting in important information is lost during such large-scale downsampling.

Qualitative Results: The qualitative comparisons are visualized in Figs. 7 and 9. The proposed TFill achieved superior visual results even under challenging conditions. In Fig. 9, we compared with CA [52], PIC [60], and CRFill [54] on Celeba-HQ dataset. Our TFill generates photorealistic high-resolution (512×512) results, even when significant semantic information is missing.

Fig. 7 shows visual results on natural images that were degraded by random masks. Here, we mainly compared

(a) Masked input (b) iGPT [8] (c) VIT [11] (d) VQ [12] Comp (e) VQ [12] Recomp (f) ICT [47] w/ *Refine* (g) TFill-*Coarse*

Figure 8. **Comparing results under different token representations.** All transformers are based on the same transformer backbone [46]. For VQGAN [12], we report completed (Comp) image and recomposed (Recomp) image. ICT [47] used two-stages networks as the original paper. TFill-*Coarse* is our model with configure $\mathbb{E}$ in Table 2, *i.e*. TFill w/o the refinement network.

Method	CelebA-HQ		FFHQ	
	LPIPS↓	FID↓	LPIPS↓	FID↓
CA [52]$_{\text{CVPR'2018}}$	0.104	9.53	0.127	8.78
PIC [60]$_{\text{CVPR'2019}}$	0.061	6.43	0.068	4.61
MEDFE [29]$_{\text{ECCV'2020}}$	0.067	7.01	-	-
$\mathbb{A}$ Traditional *Conv*	0.060	6.29	0.066	4.12
$\mathbb{B}$ + Attention in G	0.059	6.34	0.064	4.01
$\mathbb{C}$ + Restrictive *Conv*	0.056	4.68	0.060	3.87
$\mathbb{D}$ + Transformer	0.051	4.02	0.057	3.66
$\mathbb{E}$ + Masked Attention	0.050	3.92	0.057	3.63
$\mathbb{F}$ + Refine Network	**0.048**	**3.86**	**0.053**	**3.50**

Table 2. Learned Perceptual Image Patch Similarity (LPIPS) and Fréchet Inception Distance (FID) for various completion networks on center masked images. In this paper, we calculate the LPIPS and FID using all images in the corresponding test sets.

the results for semantic content completion, while visualizing the easily traditional object removal results in Appendix A.3 (Figs. A.7, A.8, and A.9). GL [17], DeepFillv2 [53], and HiFill [51], while good at object removal, failed to infer shapes needed for object completion, *e.g*. the content for animals. CRFill [54] provided plausible appearance, yet the animals' shapes are unaligned, *e.g*. malposed leg and body of the dog. Our TFill inferred the correct shapes for even heavily masked objects in ImageNet, *e.g*. the fish even with head and tail separated by a large mask. It also outperformed all previous methods on high-resolution masked images in Places2, especially for some large masked regions. More comparisons are presented in Appendix A.2 (Figs. A.4, A.5, A.6). Please zoom in to see the details.

4.3. Ablation Experiments

We ran a number of ablations to analyze the effectiveness of each component in our TFill. Results are shown in Tables 2, 3, and 4, and Figs. 8 and 9.

TFill Architecture: We first evaluated components in the redesigned image completion architecture in Table 2, which experimentally demonstrates that the new architecture considerably improves the performance. Our baseline configuration ($\mathbb{A}$) used an encoder-decoder structure derived from VQGAN [12], except here attention layers were removed in advance for a pure CNN-baseline. When combined with the powerful discriminator of StyleGANv2 [24], the performance was comparable to previous state-of-the-art CNN-based PIC [60, 61]. We first added the self-attention layer [56], *not* context mapping from the encoder [52, 60], to the decoder (Generator, G) in ($\mathbb{B}$), but the performance remained similar to baseline ($\mathbb{A}$). Interestingly, when we use the proposed *restrictive CNN* in ($\mathbb{C}$) to *embed information in the local patch*, the performance improved substantially, especially for FID (relative 20.2% on CelebA-HQ). This suggests that the input feature representation is significant for the attention layer to equally deliver all messages, as explained in Fig. 4. We then improved this new baseline by adding the transformer encoder ($\mathbb{D}$), which benefits from globally delivered messages at multiple layers. Finally, we introduced masked weights to each attention layer of the transformer ($\mathbb{E}$), improving results further.

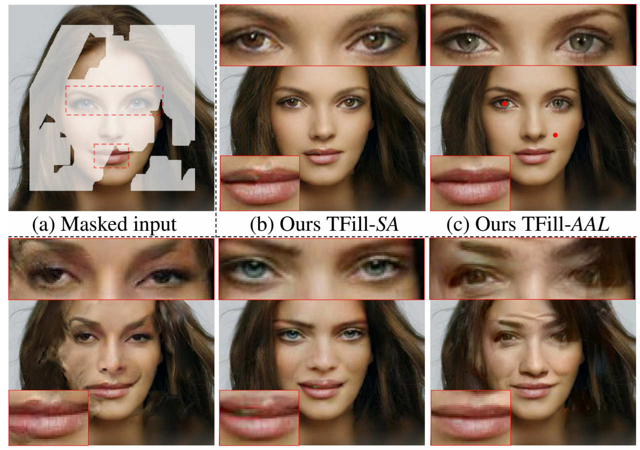
Token Representation: Tables 2 and 3 report the influence of the token representation. Our TFill achieved much better performance when using the *restrictive*-CNN. iGPT [8] downsamples the image to a fixed scale, *e.g*. 32×32, and embeds *each pixel to a token*. While this may not impact the classification [45], it has a large negative effect on generating high-quality images. Furthermore, the autoregressive form results in the completed image being inconsistent with the bottom-right visible region (Fig. 8 (b)), and each image runs an average of 26.45s on an NVIDIA 1080Ti GPU. ICT [47] improved iGPT by using bidirectional attention and adding a guided upsampling network. While the refined performance can almost match our coarse results, the running time is ruinously expensive (average 152.48s/img) and the content is *not* aligned well in Figs. 1 and 7. In contrast, VIT [11] embeds *each patch to a token*. As shown in Table 3 and Fig. 8, it can achieve relatively good quantitative and qualitative results. However, some details are perceptually poor, *e.g*. the strange eyes in Fig. 8. Finally, VQGAN [12] employs a large RF CNN to embed the image. It generates a visually realistic completion (Fig. 8 (d)), but when pasted to the original input (Fig. 8 (e)), there is an obvious gap between generated and visible pixels. When we used large convolutional kernels for large RF (229), the

Method	LPIPS↓	FID↓	Mem↓	Time↓
IGPT [8]$_{\text{ICML'2020}}$(RF 1)	0.609	148.42	3.16	26.45
VIT [11]$_{\text{ICLR'2021}}$(RF 16)	0.062	5.09	1.16	**0.167**
VQGAN [12]$_{\text{CVPR'2021}}$	0.226	11.92	2.36	4.29
ICT [47]$_{\text{ICCV'2021}}$(RF 1)	0.061	4.24	3.87	152.48
$\mathbb{E}$ TFill-*Coarse* (RF 229)	0.062	3.92	1.25	0.188
$\mathbb{E}$ TFill-*Coarse* (RF 16)	**0.057**	**3.63**	**1.15**	0.180

Table 3. The effect of restrictive token representation on FFHQ dataset. "RF" indicates the Receptive Field size. "Mem" denotes the memory (GB) cost during testing and "Time" is the testing time (s) for each center masked image.

	LPIPS↓		FID↓	
Mask Type	center	random	center	random
SA [56]$_{\text{ICML'19}}$	0.0584	0.0469	3.62	2.69
CA [52]$_{\text{CVPR'2018}}$	0.0608	0.0443	3.86	2.66
SLTA [60]$_{\text{CVPR'2019}}$	0.0561	0.0452	3.61	2.64
Ours-AAL	**0.0533**	**0.0412**	**3.50**	**2.57**

Table 4. The effect of various attention layers on FFHQ dataset. "center" denotes the center mask, "random" denotes the random mask. These attention layers were implemented within our refinement framework, while using the same content generator.



(a) Masked input (b) Ours TFill-*SA* (c) Ours TFill-*AAL*
(d) CA [52]$_{\text{CVPR'2018}}$ (e) PIC [60]$_{\text{CVPR'2019}}$ (f) CRFill [54]$_{\text{ICCV'2021}}$

Figure 9. **Results with different attention modules** in various methods. Our attention-aware layer is able to adaptively select the features from both visible and generated content.

holes will firstly be filled in with neighboring visible pixels, resulting in worse results.

AAL *vs.* Others Context Attention Modules: An evaluation of our proposed *AAL* is shown in Table 4. For this quantitative experiment we used the same content generator (our TFill-*Coarse*), but different attention modules in the refinement network. As can be seen, even using the same content, the proposed AAL reduces LPIPS and FID scores by averaging relative 6.0% and 2.8%, over the existing works [52,56,60]. This is likely due to our AAL selects features based on the largest attention scores, using weights *dynamically mapped* during inference, instead of depending on *fixed* weights to copy features as in PIC [60].

The qualitative comparison is visualized in Fig. 9. CA [52], PIC [60], and CRFill [54] used different context attention in image completion. Here, we directly use their publicly models for visualization. As can be seen in the Fig. 9, these state-of-the-art methods cannot handle large holes. While TFill-*SA* used the good but lower-resolution (256×256) coarse content from TFill-*Coarse*, the mouth exhibits artifacts with inconsistent color. Our TFill-*AAL* (TFill-*Refined*) shows no such artifacts.

5. Conclusion and Limitation

Through our analyses and experiments, we demonstrate that correctly perceiving and propagating the visible information is significantly important for masked image completion. We experimentally demonstrate the transformer-based architecture has exciting potential for content generation, due to its capacity for effectively modeling *soft*-connections between distant image content. However, unlike recent vision transformer models that either use shallow projections or large receptive fields for token representation, our *restrictive CNN projection* provides the necessary separation between explicit *global* attention modeling and implicit *local* patch correlation that leads to substantial improvement in results. We also introduced a novel attention-aware layer that adaptively balances the attention for visible and masked regions, further improving the completed image quality.

Limitations: Although our TFill model outperformed existing state-of-the-art methods on various images that were degraded by random irregular masks, the model is still not able to reason about high-level semantic knowledge. For instance, while our TFill model provided better plausible results in the third row of Fig. 7, it directly redesigned windows based on the visible windows, without understanding the physical world, that *a door is necessary for a house*. Therefore, a full understanding and imagination of semantic content in an image still needs to be further explored.

Acknowledgements: This research was supported by Monash FIT Grant. This study was also supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from Singapore Telecommunications Limited (Singtel), through Singtel Cognitive and Artificial Intelligence Lab for Enterprises (SCALE@NTU).

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001. 1, 2

[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)*, 28:24, 2009. 1, 2

[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. ACM Press/Addison-Wesley Publishing Co., 2000. 1, 2

[4] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8):882–889, 2003. 2

[5] Chenjie Cao and Yanwei Fu. Learning a sketch tensor space for image inpainting of man-made scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14509–14518, October 2021. 2

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2, 3

[7] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12299–12310, 2021. 3

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, pages 1691–1703. PMLR, 2020. 2, 3, 4, 7, 8

[9] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004. 1, 2

[10] 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. 2, 3

[11] 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. In *International Conference on Learning Representations*, 2021. 3, 4, 7, 8

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12873–12883, 2021. 2, 3, 4, 7, 8, 22

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2, 5

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 3

[15] 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. 5

[16] Drew A Hudson and C. Lawrence Zitnick. Generative adversarial transformers. *arXiv preprint*, 2021. 3

[17] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 1, 2, 6, 7

[18] Jiaya Jia and Chi-Keung Tang. Inference of segmented color and texture description by tensor voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(6):771–786, 2004. 2

[19] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 2021. 3

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

[21] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[22] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5, 12, 14, 15, 18

[23] 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. 5, 12, 14, 15, 18

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 7, 22

[25] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2

[26] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *ICCV*, volume 1, pages 305–312, 2003. 2

[27] Liang Liao, Jing Xiao, Zheng Wang, Chia-Wen Lin, and Shin'ichi Satoh. Image inpainting guided by coherence

priors of semantics and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6539–6548, 2021. 2

[28] 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)*, September 2018. 1, 2, 4, 6, 14

[29] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, 2020. 1, 7, 21

[30] 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. 2

[31] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in transformers for video inpainting. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *International Conference on Computer Vision (ICCV)*, 2021. 3, 4

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5, 12, 14, 15, 18

[34] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[35] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 1, 2, 5

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2536–2544, 2016. 1, 2

[37] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10775–10784, 2021. 1, 2

[38] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics (TOG)*, 37(4):99, 2018. 2

[39] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. 2

[40] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2, 3

[41] 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. 5, 6, 12, 14, 16, 18, 19

[42] Yuhang Song, Chao Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and C-C Jay Kuo. Contextual-based image inpainting: Infer, match, and translate. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018. 2, 3

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

[44] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 2

[45] Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008. 7

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 3, 7, 21

[47] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4692–4701, October 2021. 1, 2, 3, 4, 5, 6, 7, 8, 15, 16, 17

[48] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Masayoshi Tomizuka, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*, 2020. 3

[49] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *arXiv preprint arXiv:2106.14881*, 2021. 2, 4

[50] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the European conference on computer vision (ECCV)*, pages 1–17, 2018. 3

[51] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7508–7517, 2020. 1, 2, 3, 4, 6, 7, 16, 17

[52] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018. 1, 2, 3, 4, 5, 6, 7, 8, 15, 16, 17, 21

[53] 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. 2, 6, 7

[54] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M. Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1, 2, 3, 6, 7, 8, 16, 17

[55] Yu Zeng, Zhe Lin, Jimei Yang, Jianming Zhang, Eli Shechtman, and Huchuan Lu. High-resolution image inpainting with iterative confidence feedback and guided upsampling. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 2

[56] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 5, 7, 8

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5, 12, 14, 17, 20

[58] 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. 2

[59] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, I Eric, Chao Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations*, 2020. 2

[60] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 3, 4, 5, 6, 7, 8, 15, 16, 17, 21

[61] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic free-form image completion. *International Journal of Computer Vision*, 129(10):2786–2805, 2021. 2, 5, 7

[62] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, 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*, pages 6881–6890, 2021. 2, 3

[63] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2018. 5, 6

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 2, 3