CATN: A Context-Aggregated Transformer Network for Blind Image Inpainting

Jiahao Wang
wjhwtt@tju.edu.cn
Tianjin Uninversity

Jinyuan Li
jinyuanli@tju.edu.cn
Tianjin Uninversity

Gang Pan
pangang@tju.edu.cn
Tianjin Uninversity

Di Sun
dsun@tust.edu.cn
Tianjin University of Science and Technology

Jiawan Zhang*
jwzhang@tju.edu.cn
Tianjin Uninversity

ABSTRACT

Blind image inpainting is a challenging task that reconstructs corrupted regions without relying on mask information. Due to the lack of mask prior, previous methods usually integrate a mask prediction network at the initial phase, followed by an inpainting backbone. However, this multi-stage generation process may lead to misalignment of features. Although there are recent end-to-end generative methods that skip the mask prediction step, they have a weak perception of the contaminated regions and still cause structural distortions. This study presents a novel mask region perception strategy for blind image inpainting by combining adversarial training with forgery detection. We further propose a context-aggregated transformer network named CATN, where aggregated contextual attention (ACA) blocks are designed to alleviate the artifacts that arise from textures in contaminated regions. In particular, ACA utilizes the self-attention mechanism to better reconstruct content based on the rich context information of adjacent tokens. Moreover, we design a high-frequency omni-dimensional dynamic convolution (HODC) based on edge feature enhancement to improve the representation of texture details. Following extensive evaluations of various datasets, the proposed model CATN achieves superior performance over existing generative models in blind image inpainting, especially in terms of quality and texture.

CCS CONCEPTS

• **Computing methodologies** → **Adversarial learning**; **Image processing**; **Neural networks**.

KEYWORDS

blind image inpainting, transformer, generative adversarial network

* corresponding authors.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Image inpainting typically relies on input masks to indicate corrupted regions, which are crucial for guiding the restoration process. However, it is difficult to acquire masking information in practical applications, leading to the poor performance of inpainting algorithms that are dependent on prior knowledge. Thus, this situation promotes the development of mask-free image restoration, commonly known as blind image inpainting.

Considering the difficulty in accurately identifying corrupted parts, blind image inpainting is categorized into two distinct methods: end-to-end generation and multi-stage generation. Given a contaminated image like Figure 1(a), end-to-end methods [3, 27, 49] usually employ general inpainting frameworks and combine with Generative Adversarial Networks (GANs) [9], transformer blocks [8, 32, 47], etc to further enhance performance. Leveraging the feature inference capability of convolutional neural networks (CNNs), these frameworks can directly fill the corrupted regions of the image without using mask information as a reference. Although the end-to-end idea simplifies the process, the lack of mask perception potentially interferes with the attention to features affected by contamination, leading to a blurred texture in the final result, as illustrated in Figure 1(d).

The multi-stage methods [23, 33–35, 39, 50] decompose blind image inpainting into two sub-tasks: mask prediction and universal image inpainting. Previous works [23, 33, 39] mainly adopt CNNs to locate visually unreasonable regions. Considering that the initial mask prediction network significantly influences the reconstructed content, Ft-tdr [34] utilizes the transformer backbone for mask prediction. TransHAE [50] applies a hybrid transformer encoder with a cross-layer dissimilarity prompt, and merges two sub-tasks into one framework. However, these methods usually lead to misaligned features between the generated mask priors and the subsequent reconstructed regions. The contextual structure distortion of the final result caused by the deviations in mask prediction is illustrated in Figure 1(b-c).

Blind image inpainting requires not just the reconstruction of coherent content and fine texture, but the perception of contaminated regions. Multi-stage methods necessitate the predicted mask to represent contaminated regions, while the continual refinement of mask prediction network tends to increase the complexity of overall framework. Although end-to-end methods offer a more streamlined solution, they essentially rely on the inherent repair capabilities of the network, and they do not contain mask region perception

(a) Input (b) VCNet [50] (c) TransHAE [39]

(d) OmniNet [27] (e) Ours (f) GT

Figure 1: Comparison samples of different methods on CelebAMask-HQ dataset. (b-c) are typical of the multi-stage generation and (d) is typical of the end-to-end generation.

process. Therefore, integrating a mask region feedback mechanism into end-to-end methods is considered as an effective solution.

In addition, it is necessary to improve the feature modeling capability of the backbone inpainting network. In recent years, employing transformers to achieve global perception has become the mainstream scheme. In practice, the self-attention matrix is based on pairs of isolated queries and keys, which inadvertently limits the ability to capture fine differences in local features due to ignoring complex contextual relationships existing between tokens located at adjacent spatial locations. However, this ability is important in scenes where the texture of contaminated regions is similar to the background regions. Furthermore, current studies [25, 38] indicate that the self-attention mechanism inherently amounts to a low-pass filter. Although recent works [19, 22, 27, 46, 51] use transformer blocks to model remote context for better image inpainting, they tend to produce low-resolution quality, which could result in texture blurring.

In this paper, we address the above issues by proposing a context-aggregated transformer network, named CATN. The key idea of CATN is to combine forgery region detection with adversarial learning in the inpainting process, which provides an innovative mask region perception strategy for end-to-end models. Specifically, the generator accurately identifies and reconstructs reasonable content in corrupted regions without mask priors. The discriminator adds pixel-level perception, which locates the inpainted regions towards the mask groundtruth from the perspective of forgery detection. Note that only the computational costs of the generator are produced during inference, which means that the discriminator has the potential to integrate more complex components and thus improve its mask region perception abilities. Moreover, we design a novel aggregated contextual attention (ACA) to improve the ability to feature reconstruction, ACA introduces parallel convolutions with various kernel sizes for context mining among the key, query, and value components, thereby integrating both the local context of adjacent tokens and non-local learning. Meanwhile, we develop a frequency-full dimensional dynamic convolution (HODC) to extract certain features from ACA blocks and sampling layers in CATN. This module extends upon omni-dimensional dynamic convolution by combining edge features, thereby amplifying the representation of details.

The main contributions are summarized as follows:

- We offer a new perspective into blind image inpainting. The combination of adversarial training with forgery region detection strengthens the perception of contaminated areas, allowing the model to synthesize the accurate contents.
- We present a context-aggregated transformer network, which can perform inpainting operations in an end-to-end manner, utilizing the proposed mask region perception strategy.
- We design an aggregated contextual attention algorithm to capture both long-range dependencies and local contextual features, thereby enhancing the capacity of reconstruction.
- We develop a high-frequency omni-dimensional dynamic convolution, which incorporates edge features to improve the representation of details.

2 RELATED WORK

2.1 Image Inpainting

Conventional image inpainting primarily relies on diffusion [2, 4] or patch-matching [1, 6] schemes, which find similar segments within the original image to fill in the corrupted parts. However, these methods fail to address distortions that involve extensive content. Subsequently, deep learning becomes a mainstream technique in the field of image inpainting. Related works [7, 12, 20, 24] utilize the encoder-decoder structure and enhance contextual appearances by the development of advanced modules, such as GAN loss [28], gated convolution [44], contextual attention [21], etc. While effective in fixing abnormal features, they have difficulty in the reconstruction of large missing regions. To capture information located far apart spatially, mainstream methods [36, 37, 43] integrate pixel-wise attention blocks into the models, primarily reinforcing global context. Recently, researchers have gradually turned towards the use of transformers [19, 22, 27, 46, 47, 51], which are suitable for non-local modeling and can effectively understand and reconstruct image content across wide spatial extents. These methods rely on the mask information for inpainting, which constrains its effectiveness in scenarios without such mask data. To address this limitation, researchers develop a new approach known as blind image inpainting that allows the recovery of corrupted regions without any mask prior.

2.2 Blind Image Inpainting

Existing blind image inpainting methods include end-to-end generation and multi-stage generation. Cai *et al.* [3] first propose blind image inpainting with an end-to-end CNN architecture, which detects and restores corrupted regions without mask reference. Following this, Zhang *et al.* [49] design a feature-oriented blind inpainting network for deep face verification. Liu *et al.* [23] introduce residual modules to synthesize the details and structures. These methods typically focus on simple patch regions. To handle

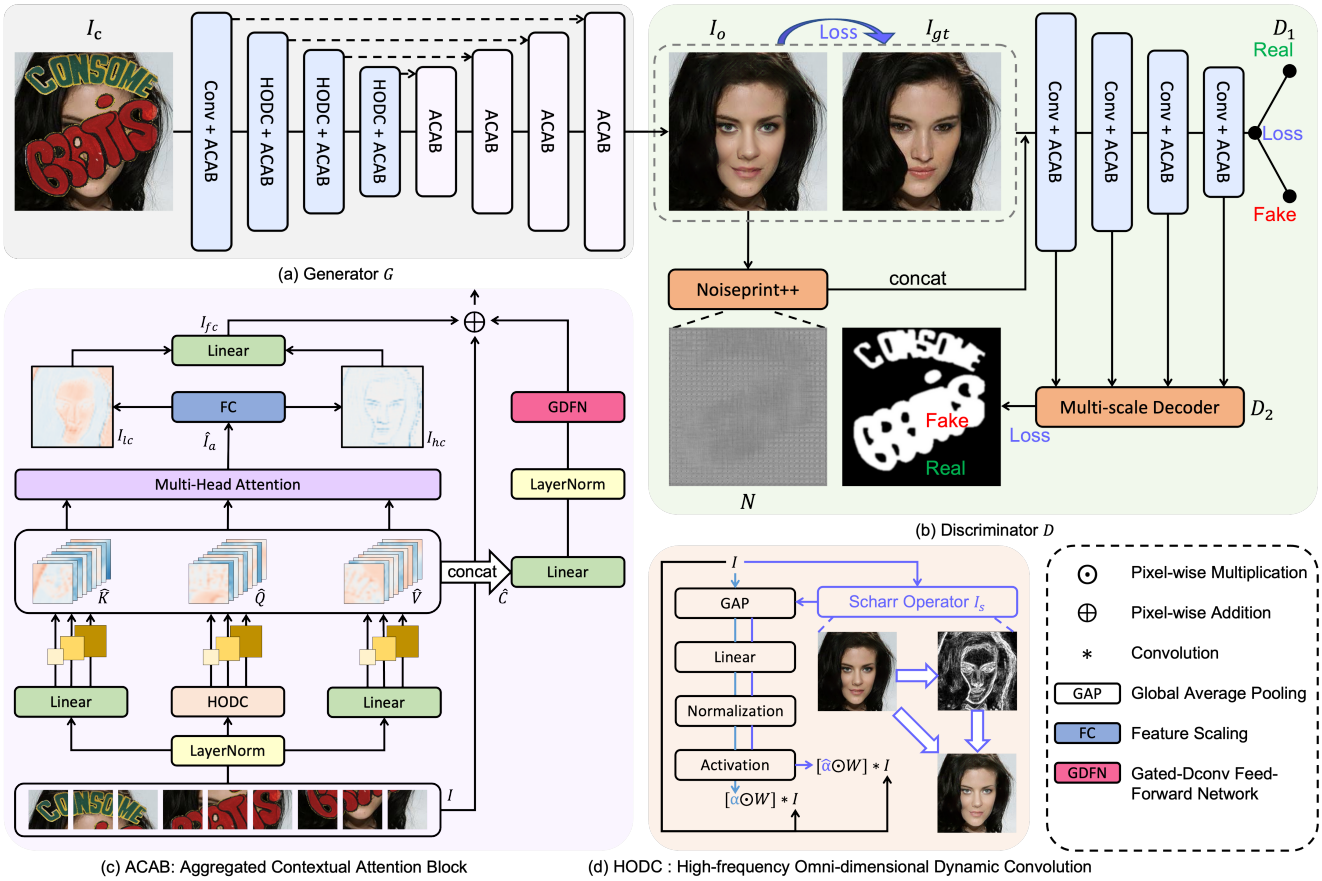


Figure 2: Framework Overview. The CATN consists of a generator G and a discriminator D. G integrates aggregated contextual attention (ACA) blocks to effectively capture both long-range dependencies and local contextual features, and the high-frequency omni-dimensional dynamic convolution (HODC) is introduced to improve texture details. D employs not just a standard binary classification mechanism D_1 for determining the overall authenticity of I_o/I_{gt}, but integrates a multi-scale decoder D_2 to perform pixel-level forgery region detection. Note that D is guided by the analysis of noise fingerprints N.

complex forms of image contamination, Wang *et al.* [39] define a two-stage framework VCN, which predicts the mask regions before inpainting, This approach accurately guides the content filling process. Similarly, SIN [35] perceives context information of the corrupted parts via self-prior learning to promote semantically coherent image synthesis. Considering the exhibit limitation when dealing with larger contaminated regions, recent works apply transformers to model long-range dependencies. For instance, Ft-tdr [34] employs self-attention blocks in both the mask prediction stage and the inpainting stage for better facial feature restoration. TransHAE [50] merges global modeling of the transformer and local modeling of CNN into a single framework to reconstruct the image. Phutke *et al.* [27] skip the mask prediction and design an end-to-end transformer-based backbone. Nevertheless, isolated interactions among keys, queries, and values in the transformer may lead to underutilized local contextual information, which tends to produce coarser structures. Additionally, a common limitation is their inadequate capacity to detect and process high-frequency patterns, often resulting in blurred texture during the learning and optimization phases. Therefore, our work aims to aggregate long-range modeling and local context representation into a transformer module while improving texture perception. The proposed framework employs a novel mask region perception strategy, which combines adversarial training with forgery detection to achieve reasonable image synthesis.

3 APPROACH

In this work, we propose an end-to-end framework named CATN (see Figure 2), which consists of a generator G and a two-branch discriminator D. Specifically, the generator G directly restores corrupted regions in the absence of mask priors. To enhance the ability for visual representation, two major components are introduced namely: (a) aggregated contextual attention (ACA), to synergistically model both global features and local contextual details, and (b) high-frequency omni-dimensional dynamic convolution (HODC): for facilitating the perception of texture information. The discriminator D focuses on improving the quality of overall appearance. Inspired by forgery region detection, the proposed CATN combines

adversarial strategies with pixel-level detection of the inpainted areas, and this advanced discriminator can be used as a mask region feedback mechanism.

Let h, w be the spatial size, $I_{gt} \in \mathbb{R}^{h\times w\times 3}$ be the groundtruth image and M be the mask image (the values 1 and 0 indicate the contaminated and uncontaminated pixels, respectively). The corrupted input image I_c is expressed as below:

$$I_c = I_{gt} \odot (1 - \mathbb{G}[M]) + N \odot \mathbb{G}[M], \tag{1}$$

where $\odot$ is pixel-wise multiplication and $N \in \mathbb{R}^{h\times w\times 3}$ is a noisy visual signal. $\mathbb{G}[\cdot]$ refers to Gaussian smoothing, a technique in image processing that employs a Gaussian filter to reduce noise and detail. This process makes image stitching smoother and even renders the contaminated areas less noticeable. The following sections will describe the framework architecture and image computation.

3.1 Framework Architecture

Generator. The generator G is an encoder-decoder network comprising 8 transformer-style ACA blocks and several sampling layers, as illustrated in Figure 2(a). The input image I_c to the encoder is initially processed into basic features by a convolution layer, then it sequentially passes through ACA blocks and downsampling layers, progressively reducing the image size (height, width) to 1/8 of its original dimensions. Conversely, the decoder employs upsampling layers and analogous processes to reconstruct the image to its original input dimensions. Meanwhile, the skip connections are added in each feature scale to retain low-level information. Given that traditional downsampling layers may result in the loss of high-frequency information, we implement HODC as a replacement to mitigate the loss of texture detail.

Discriminator. The discriminator D is structured to recover realistic details. In this study, D is also responsible for evaluating the genuineness of the restored regions. To this end, we integrate image forgery detection into the adversarial mechanism. As shown in Figure 2(b), the inpainted image I_o generated by G is processed through the Noiseprint++ [10] algorithm to extract noise-sensitive fingerprints N, which are then concatenated with I_o as input to D. The encoder of D consists of layers for downsampling and ACA blocks, whose output is divided into two branches. One branch D_1 employs binary classification for a holistic assessment of authenticity, assigning a value of 1 for real and 0 for fake. The other branch D_2 leads to a multi-scale decoder that produces pixel-level labeling maps, identifying forged areas as fake and genuine areas as real.

3.2 High-frequency Omni-dimensional Dynamic Convolution

Based on the Nyquist-Shannon sampling theorem [29], the information lost in the process of downscaling is primarily high-frequency content [40], and this challenge is compounded by the low-pass filtering nature of transformers. To maximize the retention of texture, we propose a high-frequency omni-dimensional dynamic convolution (HODC) illustrated in Figure 2(d) (the purple path), which is adaptable for replacing conventional convolutional layers such as downsampling layers and embedding layers in self-attention.

Typically, dynamic convolution [5] selects n convolutional kernels W based on the input data, rather than using a single kernel

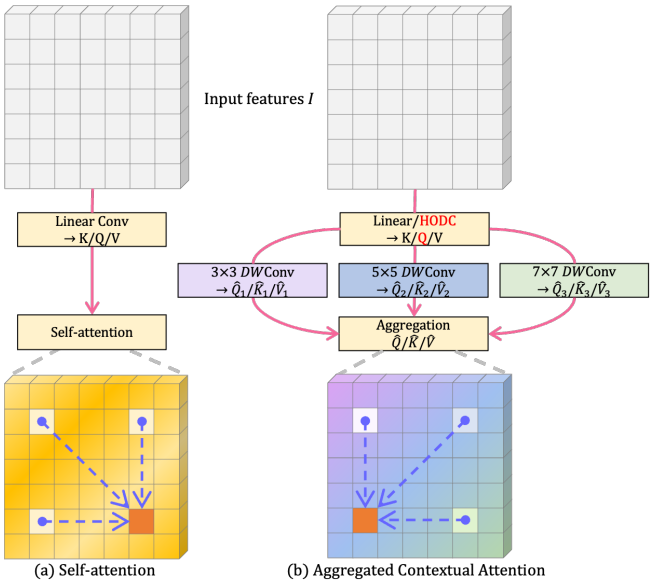


Figure 3: Visual comparison between self-attention and aggregated contextual attention (ACA). In (a), self-attention computes correlations among individual tokens, each representing a discrete element within the spatial domain. In (b), ACA adds parallel local convolutions with kernels of 3x3, 5x5, 7x7. The aggregated keys/queries/values perform attention mapping, and each token contains its neighboring contextual information. Note that the queries are generated by the high-frequency omni-dimensional convolution (HODC) instead of the linear layer.

in standard convolution. Later, the omni-dimensional dynamic convolution (ODC) [18] simultaneously selects four key dimensions of input features that specifically pertain to spatial ($\alpha_s \in \mathbb{R}^{k\times k}$, k is the kernel size), channel ($\alpha_c \in \mathbb{R}^{c_{in}}$), filter ($\alpha_f \in \mathbb{R}^{c_{out}}$), and kernel ($\alpha_w \in \mathbb{R}$). Figure 2(d) (the blue path) shows that the convolutional sets $\alpha = [\alpha_s, \alpha_c, \alpha_f, \alpha_w]$ are generated through a series of attention processes $\mathbb{P}[\cdot]$, which include global average pooling (GAP), linear projection, normalization, and Softmax/Sigmoid calculation. Given the features $I \in \mathbb{R}^{d\times c_{in}}$ (d is spatial size and c is channel) from intermediate layers of CATN, the ODC scheme can be formulated as:

$$\begin{aligned} \alpha_s, \alpha_c, \alpha_f, \alpha_w &= \mathbb{P}[I], \\ I_{odc} &= \sum_{i=1}^{n} (\alpha_{w_i} \odot \alpha_{f_i} \odot \alpha_{c_i} \odot \alpha_{s_i} \odot W_i) * I, \end{aligned} \tag{2}$$

where $I_{odc} \in \mathbb{R}^{d\times c_{out}}$ is the output features, $*$ is the convolution operation.

To amplify the representation of details, HODC employs images created through edge detection (e.g., Scharr filter) to augment the fine details in the input features. Specifically, the Scharr operator computes the gradients of I at each point in the horizontal and vertical directions. This process is achieved by performing convolution

(a) Input (b) VCNet [50] (c) TransHAE [39] (d) MAT [19] (e) OmniNet [27] (f) Ours(~~D~~) (g) Ours (h) GT

Figure 4: Comparison with the state-of-the-art. These images come from CelebAMask-HQ [16], FFHQ [14] with various contamination patterns. Ours(~~D~~) refers to the configuration where the CATN model is trained without employing the proposed mask region perception strategy denoted as *D*.

with the Scharr kernels W_x and W_y, respectively:

$$W_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}, \quad W_y = \begin{bmatrix} 3 & 10 & 3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix}. \tag{3}$$

We aim to enhance the input features with the detected edge details, so the magnitude of the gradient feature S at each point and its weighted sum $I_s \in \mathbb{R}^{d \times c_{in}}$ with the original features is computed as follows:

$$\begin{aligned} S &= \sqrt{(W_x * I)^2 + (W_y * I)^2}, \\ I_s &= \beta_1 I + \beta_2 S, \end{aligned} \tag{4}$$

where β_1, β_2 are weights that control the contribution of the original image and the edge detail. In this work, we set $\beta_1 = 1$ and $\beta_2 = 0.5$, which means the enhanced image retains the original colors and brightness while emphasizing the texture. Finally, the output feature $I_{hodc} \in \mathbb{R}^{d \times c_{out}}$ can be represented as:

$$\begin{aligned} \hat{\alpha}_s, \hat{\alpha}_c, \hat{\alpha}_f, \hat{\alpha}_w &= \mathbb{P}\left[I_s\right], \\ I_{hodc} &= \sum_{i=1}^{n} (\hat{\alpha}_{w_i} \odot \hat{\alpha}_{f_i} \odot \hat{\alpha}_{c_i} \odot \hat{\alpha}_{s_i} \odot W_i) * I. \end{aligned} \tag{5}$$

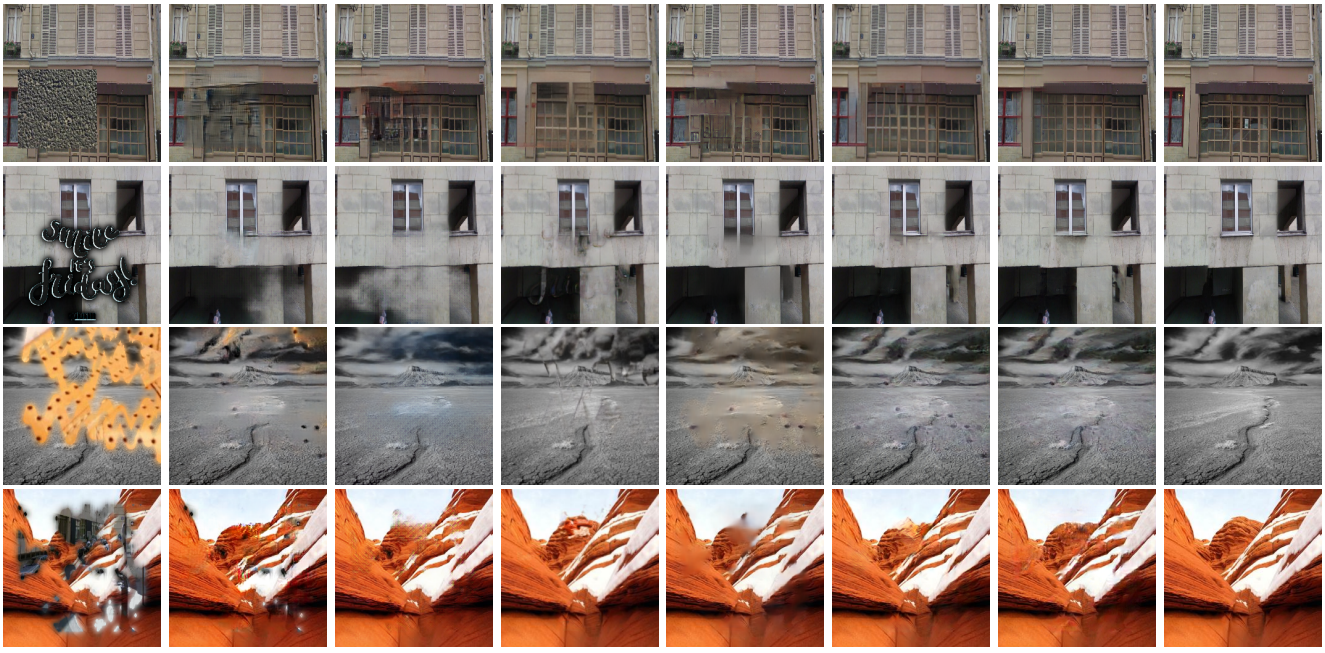
Compared to those [17, 45] that rely on additional high-frequency gating components to modulate the output features, the HODC enhances the propensity of sampling fine texture details by adaptively selecting the omni-dimensional attention α for various convolutional kernels.

3.3 Aggregated Contextual Attention

The self-attention mechanism focuses on the correlations between pairs of individual tokens. For the features $I \in \mathbb{R}^{d \times c}$, the attention first converts I into queries Q, keys K, and values V using respective linear matrices , and the output $I_a \in \mathbb{R}^{d \times c}$ is expressed as:

$$I_a = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V. \tag{6}$$

Note that all the pairwise query-key relations are independently learned over isolated query-key pairs. However, treating each query-key pair as an independent unit ignores the complex contextual relationships between tokens, which restricts the capacity to capture the nuanced distinctions within local features. To address this, we develop a novel scheme named aggregated contextual attention (ACA), which integrates local context computation with the transformer, as illustrated in Figure 2(c). Specifically, the input features I are similarly processed through linear layers to produce K and V components, while Q is generated using a HODC layer. Subsequently, each type of these components is further segmented into three parts, resulting in distinct sets $Q_i, K_i, V_i \in \mathbb{R}^{d \times c/3}, i = [1, 2, 3]$. As shown in Figure 3, different from conventional self-attention, the depth-wise convolutions (DWConv) with various kernel sizes are introduced as aggregators to collect local contextual information,



(a) Input (b) VCNet [50] (c) TransHAE [39] (d) MAT [19] (e) OmniNet [27] (f) Ours(~~D~~) (g) Ours (h) GT

Figure 5: Comparison with the state-of-the-art. These images come from Paris StreetView [26] and Places2 [52] with various contamination patterns. Ours(~~D~~) refers to the configuration where the CATN model is trained without employing the proposed mask region perception strategy denoted as *D*.

and the output features can be expressed as:

$$\begin{aligned}\hat{Q}_1, \hat{K}_1, \hat{V}_1 &= \text{DWConv}_{3\times3}(Q_1, K_1, V_1),\\ \hat{Q}_2, \hat{K}_2, \hat{V}_2 &= \text{DWConv}_{5\times5}(Q_2, K_2, V_2),\\ \hat{Q}_3, \hat{K}_3, \hat{V}_3 &= \text{DWConv}_{7\times7}(Q_3, K_3, V_3).\end{aligned} \tag{7}$$

After concatenating the same type of components from three sets, we obtain the aggregated $\hat{Q}, \hat{K}, \hat{V} \in \mathbb{R}^{d\times c}$. This process efficiently utilizes the contextual information among neighboring tokens to enhance self-attention learning. To minimize memory usage and computational demands, the subsequent step utilizes dot-product multi-head attention (MHA), as described in [30], and the output $\hat{I}_a \in \mathbb{R}^{d\times c}$ can be formulated as:

$$\hat{I}_a = \frac{\hat{Q}}{\sqrt{d}}\text{softmax}(\hat{K}^T\hat{V}). \tag{8}$$

During the aggregation process, we conduct local context convolutions for K, Q, V without diminishing the spatial resolution, thereby enhancing the detailed representation of global features. However, recent studies [25, 38] have mathematically shown that these self-attention blocks actually act as a series of low-pass filters. This common practice in transformer architecture tends to suppress high-frequency information, which can inadvertently result in the loss of textural nuances and finer details. Inspired by [38], we adopt feature scaling (FC), which dissects the output signals from MHA into the low-frequency components (LC) and high-frequency components (HC). Following this decomposition, it utilizes two distinct sets of parameters to selectively adjust the weight of these two components for each channel. Let $\mathbb{A}[\cdot]$ be the column average of a matrix, the scaled feature $I_{fc} \in \mathbb{R}^{d\times c}$ be calculated as follows:

$$\begin{aligned}I_{lc} &= \text{diag}(\gamma_1)\cdot\mathbb{A}\left[\hat{I}_a\right],\\ I_{hc} &= \text{diag}(\gamma_2)\cdot(\hat{I}_a - \mathbb{A}\left[\hat{I}_a\right]),\\ I_{fc} &= \boldsymbol{I}_{lc} + \boldsymbol{I}_{hc},\end{aligned} \tag{9}$$

where $\gamma_1, \gamma_2 \in \mathbb{R}^c$ are learnable parameters to perform channel-wise re-weighting.

Formally, a complete transformer includes a feed-forward network. Considering the application in image synthesis, ACA employs the gated-dconv feed-forward network (GDFN) from Restormer [47] to guide high-quality output. Additionally, to fully utilize the aggregated features, the ACA block adopts a parallel configuration of MHA and GDFN sub-blocks, which can be expressed as:

$$I_{aca} = I + \text{Linear}(I_{fc}) + \text{GDFN}(\sigma(\text{Linear}(\hat{C}))), \tag{10}$$

where σ is LayerNorm and $\hat{C}$ is the concatenation of $\hat{Q}, \hat{K}, \hat{V}$.

3.4 Loss Function

Taking into account the consistency between overall content and fine detail, CATN applies four types of loss functions: mean squared error (MSE) loss, perceptual loss, stochastic structural similarity (S3IM) loss [41], and GAN loss.

Content Loss. The generator G is designed to take a corrupted image I_c as input and aims to reconstruct the output image I_o towards the groundtruth image I_{gt}. The formulation of this loss

Table 1: Quantitative evaluations on the CelebAMask-HQ [16], FFHQ [14], Paris StreetView [26] and Places2 [52] with various contamination patterns as input. ↓ indicates the lower the better while ↑ means the higher the better. Ours($\not{D}$) refers to the configuration where the CATN is trained without employing the proposed mask region perception strategy denoted as D.

	Dataset	VCNet [50]	TransHAE [39]	MAT [19]	OmniNet [27]	Ours($\not{D}$)	Ours
PSNR ↑	CelebAMask-HQ	24.4288	27.3579	26.5847	24.8500	27.4748	**28.2603**
	FFHQ	23.2432	26.9964	25.7812	23.1101	26.9736	**27.1040**
	Paris StreetView	23.7850	24.9231	25.0484	22.8219	26.7190	**26.9927**
	Places2	25.0681	25.4577	26.0403	24.8325	25.9581	**26.7409**
SSIM ↑	CelebAMask-HQ	0.8871	0.9005	0.9157	0.8997	0.9263	**0.9387**
	FFHQ	0.8988	**0.9163**	0.9112	0.9010	0.9087	0.9124
	Paris StreetView	0.8275	0.8626	0.8713	0.8025	0.8700	**0.8724**
	Places2	0.8615	0.8882	0.8741	0.8291	0.8857	**0.8983**
$\ell_1(\%)$ ↓	CelebAMask-HQ	4.3712	2.6468	3.8901	4.9374	2.0953	**1.8316**
	FFHQ	4.2832	**2.0420**	3.7285	5.0538	2.2184	2.1642
	Paris StreetView	5.8475	3.1092	3.8565	4.4269	2.9211	**2.8544**
	Places2	4.7277	3.0596	2.3656	3.8230	2.2637	**2.1702**
LPIPS ↓	CelebAMask-HQ	0.1380	0.0722	0.0651	0.1424	0.0486	**0.0411**
	FFHQ	0.1125	0.0866	0.0874	0.1840	0.0504	**0.0459**
	Paris StreetView	0.1653	0.0991	**0.0795**	0.2173	0.0813	0.0805
	Places2	0.0921	0.0941	0.0825	0.1480	0.0842	**0.0722**
FID ↓	CelebAMask-HQ	13.2764	11.9616	10.9558	14.9926	10.6353	**8.4829**
	FFHQ	13.3812	11.6033	12.0317	15.2021	11.3538	**10.3784**
	Paris StreetView	52.1438	35.8904	38.3674	43.4504	36.3674	**34.9745**
	Places2	27.2467	23.4471	24.4273	23.1912	22.3898	**20.5134**

function is as follows:

$$\mathcal{L}_{con} = \|I_o \odot I_{gt}\|_2^2, \tag{11}$$

where $\odot$ is the pixel-wise multiplication and $\|\cdot\|_2$ is the Euclidean norm.

Perceptual Loss. To improve the perceptual quality of images, we adopt a perceptual loss function using a pre-trained VGG-16 network [31].

$$\mathcal{L}_{perc} = \sum_i \left\|\Phi_i\left(I_o\right) - \Phi_i\left(I_{gt}\right)\right\|_1, \tag{12}$$

where Φ_i represents the output feature map of the i-th layer in VGG-16, corresponding to the activation layers: $ReLu1\_1$, $ReLu2\_1$, $ReLu3\_1$, $ReLu4\_1$, and $ReLu5\_1$.

S3IM Loss. The structural similarity (SSIM) can capture local information from adjacent pixels through convolutional kernels, but it is limited in detecting the structural information in more distant pixel arrangements. To address this, we introduce an improved S3IM loss for inpainting:

$$\mathcal{L}_{s3im} = 1 - \text{S3IM}(I_o, I_{gt}). \tag{13}$$

Unlike processing the minibatch images in [41], our S3IM loss involves randomly shuffling the pixels of a single image I_o (including the groundtruth). This process generates non-local sets, which are subsequently calculated by the SSIM.

GAN Loss. For the overall image discrimination, we utilize the hinge loss function [13] to optimize both the projected discriminator D and the generator G. Note that D receives the concatenation of I_o and its noise fingerprints N as input:

$$\begin{aligned} N_o &= \text{Concat}(I_o, N), \\ N_{gt} &= \text{Concat}(I_{gt}, N). \end{aligned} \tag{14}$$

Thus the objective function for GAN process is formulated as follows:

$$\begin{aligned} \mathcal{L}_{adv}^D &= \mathbb{E}_{I_{gt}}[ReLU(1 - D_1(N_{gt}))] + \mathbb{E}_{I_o}[ReLU(1 + D_1(N_o))], \\ \mathcal{L}_{adv}^G &= -\mathbb{E}_{I_o}[D_1(N_o)]. \end{aligned} \tag{15}$$

Additionally, for mask region perception, we implement forgery discrimination devised to distinguish between authentic and forged pixels within an image.

$$\begin{aligned} \mathcal{L}_s^D &= \mathbb{E}_{I_{gt}}[ReLU(1 - D_2(N_{gt}))] \\ &\quad + \mathbb{E}_{I_o}[ReLU(1 - D_2(N_o) \odot (1 - M)] \\ &\quad + \mathbb{E}_{I_o}[ReLU(1 + D_2(N_o) \odot M)], \\ \mathcal{L}_s^G &= -\mathbb{E}_{I_o}[D_2(N_o) \odot M]. \end{aligned} \tag{16}$$

Total Loss. The whole loss function can be obtained as:

$$\mathcal{L} = \min_G \max_D (\mathcal{L}_{con} + \lambda_1 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{s3im} + \lambda_3 \mathcal{L}_{adv} + \lambda_4 \mathcal{L}_s) \tag{17}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyper-parameters. In this work, we empirically set $\lambda_1 = 100$, $\lambda_2 = 1$, $\lambda_3 = \lambda_4 = 0.1$.

4 EXPERIMENTS

4.1 Implementation Details

The CATN is evaluated using four public datasets including a range of subjects: CelebAMask-HQ [16] and FFHQ [14] for high-quality faces, Paris StreetView [26] and Places2 [52] for scenes. In terms of data preprocessing, all input images are contaminated by constant values, patches of the scene images, and texture images. As shown in Figure 7, we apply two contamination patterns: regular patterns

(a) Input (b) BN (c) BN+HDL (d) BN+HDL+HQ (e) BN+HDL+HQ+FS (f) BN+HDL+HQ+FS+CA (g) GT

Figure 6: Ablation study on different configurations of the CATN for blind image inpainting. The experiment is conducted on the CelebAMask-HQ [16] dataset with regular contamination patterns.

(a) GT (b) Contaminant (c) Regular (d) Irregular (e) Text-like

Figure 7: A groundtruth image can be subjected to contamination using three distinct types of patterns.

and irregular patterns (including text-like patterns [42]), to simulate various types of blind images.

During the training phase, we use the Adam optimizer [15] with hyperparameters β_1 set to 0.5 and β_2 to 0.9. The learning rate for both the generator and discriminator is configured at 1e-4. The CATN is developed using PyTorch and is trained on NVIDIA RTX 3090 GPUs.

4.2 Quantitative Evaluation

In the evaluation of inpainting results with various contamination patterns, CATN is compared with state-of-the-art such as VCNet [50], TransHAE [39], and OmniNet [27] for blind image inpainting. Meanwhile, a non-blind image inpainting method MAT [19] is applied as a comparative reference. These comparisons are conducted on testing datasets from CelebAMask-HQ, FFHQ, Places2, and Paris StreetView. Consistent with standard practices in image inpainting research, we employ Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM), and Mean ℓ_1 error as quantitative metrics, which are calculated on the spatial images to assess the accuracy of the inpainting. In addition, two additional metrics: Learned Perceptual Image Patch Similarity (LPIPS) [48] and the Frechet Inception Score (FID) [11], are utilized to measure the perceptual quality of predicted images compared to the groundtruth images. As detailed in Table 1, comparative experiments conducted on different datasets show that the proposed method outperforms existing approaches on most of the metrics.

4.3 Qualitative Evaluations

To validate the inpainting performance, Figure 4 and Figure 5 present a comparative analysis of the predicted results from different methods. As illustrated in Figure 4, the inpainting result from VCNet seems to produce distorted structures, particularly noticeable around contaminated edge regions. TransHAE tends to produce

Table 2: Ablation study on the CelebAMask-HQ dataset with regular contamination pattern.

Methods	PSNR ↑	SSIM ↑	ℓ_1(%) ↓	LPIPS ↓	FID ↓
BN	25.65	0.862	4.18	0.085	19.26
BN+HDL	25.84	0.874	3.93	0.082	17.48
BN+HDL+HQ	26.13	0.891	3.47	0.079	15.27
BN+HDL +HQ+FS	26.47	0.909	3.34	0.075	14.51
BN+HDL +HQ+FS+CA	**26.91**	**0.921**	**2.97**	**0.066**	**12.45**

texture noise during the reconstruction of features. Although MAT utilizes mask information as part of its input for non-blind image inpainting, the output still exhibits artifacts that are affected by contaminants present in the original image. OmniNet is capable of recovering reasonable content but often ignores texture details. In contrast, our method enhances the perception of contaminated regions via an adversarial training strategy to achieve accurate reconstruction. Moreover, Figure 5 shows similar results on the testing datasets. Both VCNet and TransHAE struggle with maintaining reasonable semantics and detail accuracy. While MAT and OmniNet attempt to generate plausible structures, their outputs often contain confusing artifacts. In contrast, our method produces more reliable and high-quality inpainting results.

4.4 Ablation study

Figure 4 and Figure 5 (see columns f and g) have shown that the proposed mask region perception strategy D significantly reduces the presence of contaminant artifacts. In this subsection, we continue to analyze how the other proposed modules (ACA block, HODC) contribute to the final performance of image inpainting. Specifically, we verify the effectiveness of the CATN backbone network (BN) by removing HODC and replacing ACA blocks with standard transformer blocks. Following this, the HODC downsampling layers (HDL), HODC queries (HQ), feature scaling (FS), and contextual aggregation (CA) are integrated into the backbone in turn, and this operation allows us to assess the contributions of these components to the overall performance. Figure 6 shows that these components facilitate reasonable contextual content and fine texture details on the CelebAMask-HQ dataset. Note that this dataset adopts regular contamination patterns, which are referred to as unseen patterns in TransHAE. Moreover, Table 2 illustrates that our proposed modules

demonstrably enhance the performance in the task of blind image inpainting.

5 CONCLUSION

This paper presents CATN, a robust blind inpainting framework that exhibits significant restoration ability across a range of benchmark datasets. We adopt an adversarial training strategy that integrates forgery detection as mask region perception. To expand the receptive field and simultaneously address local content features, CATN introduces aggregated contextual attention blocks. Additionally, high-frequency omni-dimensional dynamic convolution is implemented to capture more texture details. Comprehensive testing on various benchmark datasets demonstrates that CATN achieves superior results in blind image inpainting for various contamination. The proposed CATN effectively improves content reconstruction without mask priors and expands its applicability for more realistic scenarios. Additionally, ACA and HODC modules can offer valuable insights for future related tasks.

REFERENCES

[1] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. 2009. PatchMatch: A randomized correspondence algorithm for structural image editing. ToG 28, 3 (2009), 24.

[2] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. 2000. Image inpainting. In Siggraph. ACM, USA, 417–424.

[3] Nian Cai, Zhenghang Su, Zhineng Lin, Han Wang, Zhijing Yang, and Bingo Wing-Kuen Ling. 2017. Blind inpainting using the fully convolutional neural network. The Visual Computer 33 (2017), 249–261.

[4] Tony F Chan and Jianhong Shen. 2001. Nontexture inpainting by curvature-driven diffusions. JVCIR 12, 4 (2001), 436–449.

[5] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. 2020. Dynamic convolution: Attention over convolution kernels. In CVPR. 11030–11039.

[6] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. 2012. Image melding: Combining inconsistent images using patch-based synthesis. ToG 31, 4 (2012), 1–10.

[7] Ruohan Gao and Kristen Grauman. 2017. On-demand learning for deep image restoration. In ICCV. IEEE Computer Society, Italy, 1086–1095.

[8] Chongjian Ge, Xiaohan Ding, Zhan Tong, Li Yuan, Jiangliu Wang, Yibing Song, and Ping Luo. 2023. Advancing Vision Transformers with Group-Mix Attention. arXiv preprint arXiv:2311.15157 (2023).

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. 2014. Generative Adversarial Nets. In NeurIPS. Curran Associates, USA.

[10] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In CVPR. 20606–20615.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017).

[12] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. ToG 36, 4 (2017), 1–14.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In CVPR. IEEE Computer Society, USA, 1125–1134.

[14] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In CVPR. 4401–4410.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In CVPR. IEEE Computer Society, USA, 5549–5558.

[17] Bin Li, Bowei Zheng, Haodong Li, and Yanran Li. 2021. Detail-enhanced image inpainting based on discrete wavelet transforms. Signal Processing 189 (2021), 108278.

[18] Chao Li, Aojun Zhou, and Anbang Yao. 2022. Omni-dimensional dynamic convolution. arXiv preprint arXiv:2209.07947 (2022).

[19] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. 2022. Mat: Mask-aware transformer for large hole image inpainting. In CVPR. 10758–10768.

[20] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. 2018. Image inpainting for irregular holes using partial convolutions. In ECCV. Springer, USA, 85–100.

[21] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. 2019. Coherent semantic attention for image inpainting. In ICCV. 4170–4179.

[22] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. 2022. Reduce Information Loss in Transformers for Pluralistic Image Inpainting. In CVPR. IEEE Computer Society, USA, 11347–11357.

[23] Yang Liu, Jinshan Pan, and Zhixun Su. 2019. Deep blind image inpainting. In IScIDE. Springer, 128–141.

[24] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. 2019. Edgeconnect: Structure guided image inpainting using edge prediction. In ICCV Workshops. IEEE Computer Society, Korea, 0–0.

[25] Namuk Park and Songkuk Kim. 2022. How Do Vision Transformers Work? arXiv preprint arXiv:2202.06709 (2022).

[26] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In CVPR. IEEE Computer Society, USA, 2536–2544.

[27] Shruti S Phutke, Ashutosh Kulkarni, Santosh Kumar Vipparthi, and Subrahmanyam Murala. 2023. Blind Image Inpainting via Omni-Dimensional Gated Attention and Wavelet Queries. In CVPR. 1251–1260.

[28] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015).

[29] Claude E Shannon. 1949. Communication in the presence of noise. IRE 37, 1 (1949), 10–21.

[30] Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. 2021. Efficient attention: Attention with linear complexities. In WACV. 3531–3539.

[31] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NeurIPS, Vol. 30. Curran Associates, USA.

[33] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. 2020. Bringing old photos back to life. In CVPR. 2747–2757.

[34] Junke Wang, Shaoxiang Chen, Zuxuan Wu, and Yu-Gang Jiang. 2022. Ft-tdr: Frequency-guided transformer and top-down refinement network for blind face inpainting. IEEE TMM (2022).

[35] Juan Wang, Chunfeng Yuan, Bing Li, Ying Deng, Weiming Hu, and Stephen Maybank. 2023. Self-Prior Guided Pixel Adversarial Networks for Blind Image Inpainting. IEEE PAMI (2023).

[36] Ning Wang, Jingyuan Li, Lefei Zhang, and Bo Du. 2019. MUSICAL: Multi-Scale Image Contextual Attention Learning for Inpainting.. In IJCAI. AAAI Press, China, 3748–3754.

[37] Ning Wang, Sihan Ma, Jingyuan Li, Yipeng Zhang, and Lefei Zhang. 2020. Multi-stage attention network for image inpainting. Pattern Recognition 106 (2020), 107448.

[38] Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. 2022. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. arXiv preprint arXiv:2203.05962 (2022).

[39] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. 2020. Vcnet: A robust approach to blind image inpainting. In ECCV. Springer, 752–768.

[40] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. 2020. Invertible image rescaling. In ECCV. Springer, UK, 126–144.

[41] Zeke Xie, Xindi Yang, Yujie Yang, Qi Sun, Yixiang Jiang, Haoran Wang, Yunfeng Cai, and Mingming Sun. 2023. S3im: Stochastic structural similarity and its unreasonable effectiveness for neural fields. In ICCV. 18024–18034.

[42] Xingqian Xu, Zhifei Zhang, Zhaowen Wang, Brian Price, Zhonghao Wang, and Humphrey Shi. 2021. Rethinking text segmentation: A novel dataset and a text-specific refinement approach. In CVPR. 12045–12055.

[43] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2018. Generative image inpainting with contextual attention. In CVPR. IEEE Computer Society, USA, 5505–5514.

[44] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. 2019. Free-form image inpainting with gated convolution. In ICCV. IEEE Computer Society, Korea, 4471–4480.

[45] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. 2021. Wavefill: A wavelet-based generation network for image inpainting. In ICCV. IEEE Computer Society, Canada, 14114–14123.

[46] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. 2021. Diverse image inpainting with bidirectional and autoregressive transformers. In ACM MM. 69–78.

[47] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2022. Restormer: Efficient transformer for high-resolution image restoration. In CVPR. 5728–5739.

[48] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR. 586–595.
[49] Shu Zhang, Ran He, Zhenan Sun, and Tieniu Tan. 2017. Demeshnet: Blind face inpainting for deep meshface verification. IEEE TIFS 13, 3 (2017), 637–647.
[50] Haoru Zhao, Zhaorui Gu, Bing Zheng, and Haiyong Zheng. 2022. Transcnn-hae: Transformer-cnn hybrid autoencoder for blind image inpainting. In ACM MM. 6813–6821.
[51] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung. 2022. Bridging Global Context Interactions for High-Fidelity Image Completion. In CVPR. IEEE Computer Society, USA, 11512–11522.
[52] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. TPAMI 40, 6 (2017), 1452–1464.