AFAN: An Attention-Driven Forgery Adversarial Network for Blind Image Inpainting

Jiahao Wang , Gang Pan , Di Sun , Jinyuan Li , Jiawan Zhang

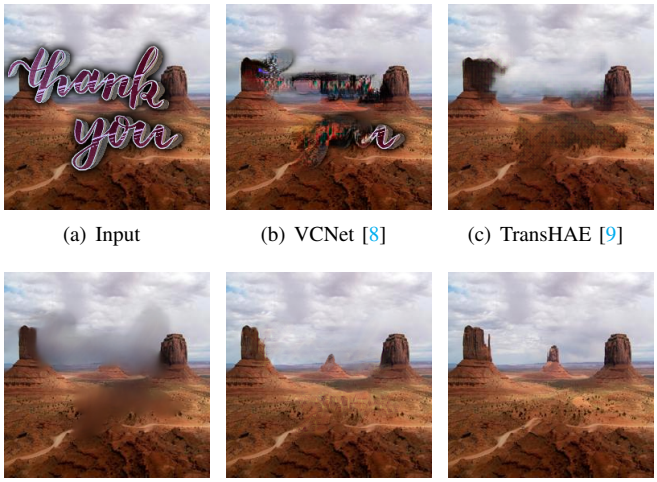
***Abstract*—Blind image inpainting is a challenging task that reconstructs corrupted regions without relying on mask information. Due to the lack of mask prior, previous methods usually integrate a mask prediction network at the initial phase, followed by an inpainting backbone. However, this multi-stage 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 an attention-driven forgery adversarial network named AFAN, which consists of adaptive contextual attention (ACA) blocks to achieve feature modulation. In particular, within the generator, ACA utilizes self-attention to better reconstruct content based on the rich context information of adjacent tokens. In the discriminator, ACA utilizes cross-attention with noise priors to guide the learning of adversarial forgery detection. Moreover, we design a high-frequency omni-dimensional dynamic convolution (HODC) based on edge feature enhancement to improve the detail representation. Following extensive evaluations of various datasets, the proposed model AFAN achieves superior performance over existing generative models in blind image inpainting, especially in terms of quality and texture.**

***Index Terms*—Blind image inpainting, transformer, generative adversarial network.**

I. 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 Fig. 1(a), end-to-end methods [1], [2], [3] usually employ general inpainting frameworks and combine with Generative Adversarial Networks (GANs) [4], transformer blocks [5], [6], [7], etc to further enhance performance. Leveraging the feature inference capability of backbone networks, 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 Fig. 1(d).



(a) Input (b) VCNet [8] (c) TransHAE [9]

(d) OmniNet [3] (e) Ours (f) GT

Fig. 1. Comparison samples of different methods on Places2 dataset. (b-c) are typical of the multi-stage generation and (d) is typical of the end-to-end generation.

The multi-stage methods [10], [11], [9], [12], [8], [13] decompose blind image inpainting into two sub-tasks: mask prediction and universal image inpainting. Previous works [10], [11], [9] mainly adopt convolutional neural networks (CNNs) to locate visually unreasonable regions. Considering that the initial mask prediction network significantly influences the reconstructed content, Ft-tdr [12] utilizes the transformer backbone for mask prediction. TransHAE [8] 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 Fig. 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

J. Wang, G. Pan, J. Li, and J. Zhang (corresponding author) are with the College of Intelligence and Computing, Tianjin University, Tianjin, 300350 China (e-mail: wjhwtt@tju.edu.cn; pangang@tju.edu.cn; jwzhang@tju.edu.cn).

D. Sun is with Tianjin University of Science and Technology, Tianjin, 300457 China (e-mail: dsun@tust.edu.cn).

rely on the inherent repair capabilities of the network, and they do not contain mask region perception process. Therefore, integrating a mask region feedback mechanism into end-to-end methods is considered as an effective solution.

In this paper, we address the above issues by proposing an attention-driven forgery adversarial network, named AFAN. The key idea of AFAN 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 and noise priors, 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.

For the feature modeling capability of the AFAN, employing transformers to achieve global perception has become the mainstream scheme. In practice, the 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 blind image scenes where the texture of contaminated regions is similar to the background regions. To address this, we design a novel adaptive contextual attention (ACA) to improve the feature modeling ability by integrating the local context of adjacent tokens with non-local learning. Specifically, within the generator, the ACA introduces a gating mechanism in the query component to dynamically fuse multi-scale features that contain local contextual information. In the discriminator, the ACA reconstructs the noise priors as the query component using the same process. This query component then engages in cross-attention with key-value pairs derived from features of the inpainted image, thereby guiding adversarial training for forgery detection. Moreover, we develop a high-frequency omni-dimensional dynamic convolution (HODC) to further modulate local details. This module extends upon omni-dimensional dynamic convolution by combining edge features, thereby highlighting the contaminated regions and amplifying the representation of texture.

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 an attention-driven forgery adversarial network, which can perform inpainting operations in an end-to-end manner, utilizing the proposed mask region perception strategy.
- We design an adaptive 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.

II. Related Work

A. Image Inpainting

Conventional image inpainting primarily relies on diffusion [14], [15] or patch-matching [16], [17] 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 [18], [19], [20], [21] utilize the encoder-decoder structure and enhance contextual appearances by the development of advanced modules, such as GAN loss [22], gated convolution [23], contextual attention [24], [25], 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 [26], [27], [28] integrate pixel-wise attention blocks into the models, primarily reinforcing global context. Recently, researchers have gradually turned towards the use of transformers [29], [30], [31], [32], [6], [33], 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.

B. Blind Image Inpainting

Existing blind image inpainting methods include end-to-end generation and multi-stage generation. Cai *et al.* [1] 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.* [2] design a feature-oriented blind inpainting network for deep face verification. Liu *et al.* [10] introduce residual modules to synthesize the details and structures. These methods typically focus on simple patch regions. To handle complex forms of image contamination, Wang *et al.* [9] define a two-stage framework VCN, which predicts the mask regions before inpainting, This approach accurately guides the content filling process. Similarly, SIN [13] 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 [12] employs self-attention blocks in both the mask prediction stage and the inpainting stage for better facial feature restoration. TransHAE [8] merges global modeling of the transformer and local modeling of CNN into a single framework to reconstruct the image. Phutke *et al.* [3] 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

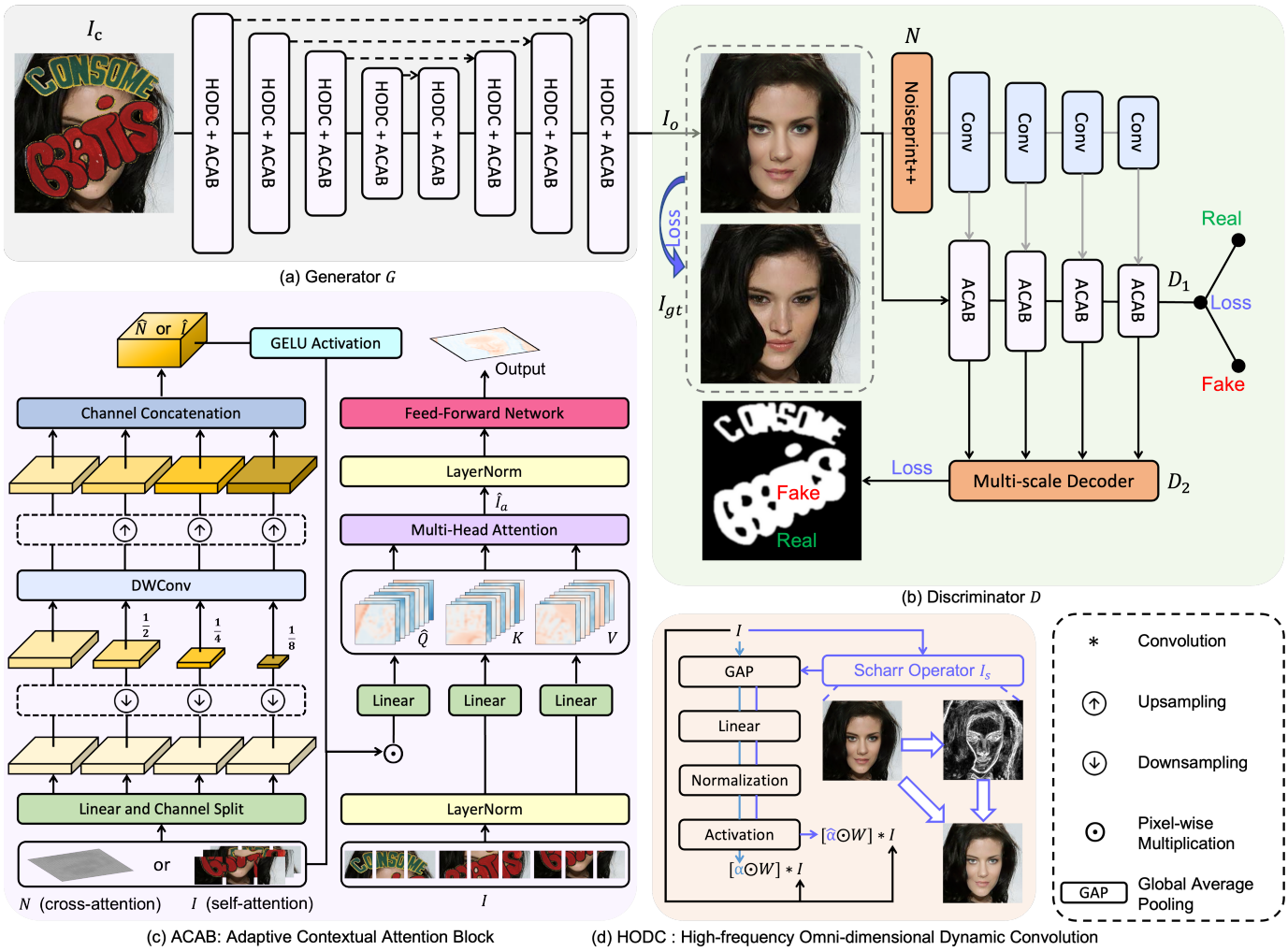


Fig. 2. Framework Overview. The AFAN consists of a generator G and a discriminator D. G integrates adaptive contextual attention (ACA) blocks to effectively capture both long-range dependencies and local contextual features, and the high-frequency omni-dimensional dynamic convolution (HODC) is introduced 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.

underutilized local contextual information, which tends to produce coarser structures. Therefore, our work aims to aggregate long-range modeling and local context representation into a transformer module. The proposed framework employs a novel mask region perception strategy, which combines adversarial training with forgery detection to achieve reasonable image synthesis.

III. Approach

In this work, we propose an end-to-end framework named AFAN (see Fig. 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) adaptive 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 AFAN 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.

(a) Input (b) Output (c) Noiseprint++ (d) Scaling

Fig. 3. Noisy fingerprint representation. (b) shows the image recovered by AFAN. (c) and (d) display the noise-sensitive fingerprints generated by the Noiseprint++ algorithm from (b) and the feature-scaled (b), respectively. Note that feature scaling in this scene refers to downsampling the image and then restoring it to its original size.

A. Adversarial Training with Forgery Detection

Motivation. For the mask-free image inpainting, the absence of mask perception could potentially weaken the restoration of contaminated regions. The proposed AFAN innovatively integrates adversarial training with forgery detection, introducing a feedback mechanism for mask regions to enhance the performance of end-to-end methods. Thus, the discrimination module is structured not just to recover realistic details but to evaluate the genuineness of the restored regions.

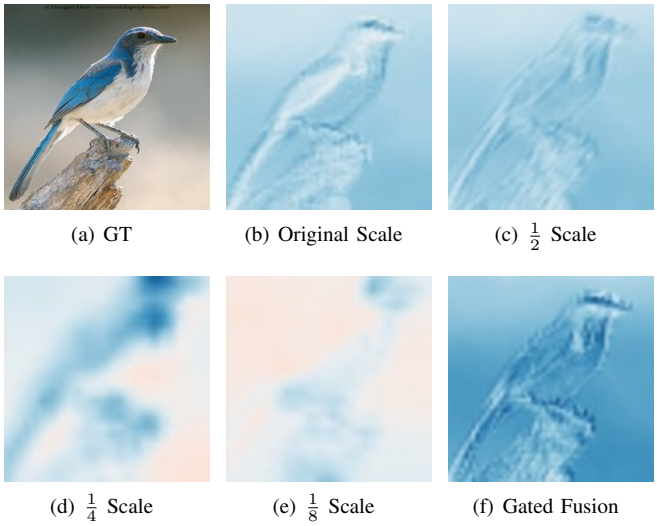
Forgery Detection. The discriminator D identifies inpainted regions from the perspective of forgery detection and aligns them with the mask groundtruth. Recent forgery detection methods usually introduce noise-sensitive fingerprints as additional input, such as Noiseprint [34], Noiseprint++ [35], and SRM filtering [36]. This work uses the state-of-the-art Noiseprint++ algorithm to generate robust noise priors N, as illustrated in Fig. 3. Even when feature scaling alters the distribution of unseen noise in the inpainted image, this algorithm effectively highlights grid inconsistencies in the edited areas (see Fig. 3(d)).

Discriminator Architecture. As shown in Fig. 2(b), the inpainted image I_o generated by the generator G is used as input to discriminator D. The encoder of D consists of layers for downsampling and ACA blocks. To enhance the robustness of forgery detection, the noise-sensitive fingerprints N are integrated into the image features using the cross-attention mechanism of the ACA blocks. Meanwhile, the output of the encoder 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 aggregates features of all downsampling stages to produce robust pixel-level labeling maps. This decoder identifies forged areas as fake and genuine areas as real.

Adversarial Training. For the overall image discrimination, this work utilizes the hinge loss function [37] to optimize both the projected discriminator D and the generator G. Thus the objective function for the GAN process is expressed as:

$$\begin{aligned}\mathcal{L}_{adv}^{D} &= \mathbb{E}_{I_{gt}}[ReLU(1 - D_1(I_{gt}, N))] \\ &\quad + \mathbb{E}_{I_o}[ReLU(1 + D_1(I_o, N))], \\ \mathcal{L}_{adv}^{G} &= -\mathbb{E}_{I_o}[D_1(I_o, N)].\end{aligned} \tag{2}$$

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



(a) GT (b) Original Scale (c) $\frac{1}{2}$ Scale

(d) $\frac{1}{4}$ Scale (e) $\frac{1}{8}$ Scale (f) Gated Fusion

Fig. 4. Multi-scale feature representation. (b)-(e) are the contextual feature representations sampled at different spatial scales in the ACA. (f) represents the gated mechanism $\mathbb{G}[\hat{I}, I]$ that adaptively fuse these multi-scale features.

$$\begin{aligned}\mathcal{L}_{s}^{D} &= \mathbb{E}_{I_{gt}}[ReLU(1 - D_2(I_{gt}, N))] \\ &\quad + \mathbb{E}_{I_o}[ReLU(1 - D_2(I_o, N) \odot (1 - M)] \\ &\quad + \mathbb{E}_{I_o}[ReLU(1 + D_2(I_o, N) \odot M)], \\ \mathcal{L}_{s}^{G} &= -\mathbb{E}_{I_o}[D_2(I_o, N) \odot M].\end{aligned} \tag{3}$$

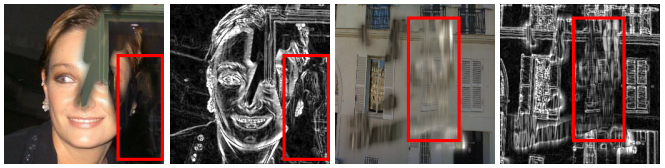
B. Generator Architecture.

As illustrated in Fig. 2(a), the generator G is an encoder-decoder network comprising 8 transformer-style components and several sampling layers. Each pair of mirrored components between the encoder and decoder contains [4, 6, 6, 8] ACA blocks, with [1, 2, 4, 8] attention heads and [48, 96, 192, 384] channels, respectively. Notably, a HODC layer is added before each block to enhance texture details, and the ACA performs self-attention instead of cross-attention in the generator. The input image I_c to the encoder sequentially passes through HODC layers (which can serve as downsampling layers) and ACA blocks, 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.

C. Adaptive Contextual Attention

The self-attention mechanism focuses on the correlations between pairs of individual tokens. Given the features $I \in \mathbb{R}^{d \times c}$ (d is spatial size and c is channel) from intermediate layers of AFAN, 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 formulated as follows:

$$I_a = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V. \tag{4}$$



(a) (b) (c) (d)

Fig. 5. Edge feature representation. (a) and (c) are the contaminated images, while (b) and (d) are the corresponding edge images obtained through Scharr filtering. The edge features amplify the global detail representation and highlight the contours and textures of the contaminated regions that are similar to the background.

Building on this, CAT [38] proposes the cross-attention that combines asymmetrically two separate embedding sequences of the same dimension.

For the discriminator D, we employ cross-attention (the noise fingerprints N serve as a query input and the inpainted image I_o as a key and value input), effectively integrating noise priors into the image features. Since the noise fingerprints N are sparse high-frequency information, applying global spatial attention to these features may be redundant and computationally expensive. Therefore, the attention operation targets the channel dimensions ($c \times c$) instead of the spatial dimension ($d \times d$):

$$I_a^D = \text{softmax}\left(\frac{Q_N^T \cdot K_I}{\sqrt{d}}\right) \cdot V_I. \tag{5}$$

Although cross-channel attention effectively recovers high-quality depth features, it lacks the compensation for spatial feature modulation. This shortfall is due to the dot product calculation treating each query-key pair as an independent unit, thus ignoring the intricate spatial contextual relationships among tokens. This limitation weakens the capacity to capture the nuanced distinctions within local features, especially for noise fingerprints N. To address this, we develop a novel scheme named adaptive contextual attention (ACA), which integrates local context computation with global attention, as illustrated in Fig. 2(c). Specifically, the noise fingerprints $N \in \mathbb{R}^{d \times c}$ are split into n parts at the channel level, resulting in a distinct set $\{N_0, N_1, ..., N_{n-1}\}, 2 \leq n \leq 4$. The first part N_0 performs depth-wise convolutions (DWConv) with kernel size $k = 2n-1$ to collect local contextual information, while the rest parts N_i ($i \in [1, n-1]$) are downsampled to $1/2^i$ of their original size through max-pooling layers. Subsequently, these multi-scale features similarly perform $k \times k$ depth-wise convolutions and restore their original size using the nearest interpolation. This process generates a new set $\{\hat{N}_0, \hat{N}_1, ..., \hat{N}_{n-1}\}$, which are then concatenated along the channel dimension to form an aggregated feature $\hat{N}$. It can be formulated as:

$$\begin{aligned} &N_0, N_1, ..., N_{n-1} = \text{Split}(N), \\ &\hat{N}_0 = \text{DWConv}_{k\times k}(N_0), \\ &\hat{N}_i = \uparrow_{2^i} (\text{DWConv}_{k\times k}(\downarrow_{\frac{1}{2^i}} (N_i))), \\ &\hat{N} = \text{Concat}(\hat{N}_0, \hat{N}_1, ..., \hat{N}_{n-1}), \end{aligned} \tag{6}$$

where $\downarrow$ and $\uparrow$ represent the downsampling and upsampling operations, respectively. The feature $\hat{N}$ contains rich spatial context, which can enhance the detailed representation of the initial feature N. To this end, we apply a gated mechanism $\mathbb{G}[\cdot]$ to adaptively fuse them:

$$\mathbb{G}\left[\hat{N}, N\right] = \phi(\hat{N}) \odot N, \tag{7}$$

where ϕ is GELU activation function and $\odot$ is pixel-wise multiplication. Meanwhile, a new $\hat{Q}$ component is generated based on the fused features, and the output $\hat{I}_a^D \in \mathbb{R}^{d \times c}$ of ACA is calculated as follows:

$$\begin{aligned} &\hat{Q}_N = \text{Conv}_{1\times1}(\mathbb{G}\left[\hat{N}, N\right]), \\ &\hat{I}_a^D = \text{softmax}\left(\frac{\hat{Q}_N^T \cdot K_I}{\sqrt{d}}\right) \cdot V_I, \end{aligned} \tag{8}$$

here the K_I and V_I components are derived from the inpainted image features I_o via linear layers. This scheme efficiently utilizes the contextual information among neighboring tokens to enhance non-local learning.

For the generator G, the enhancement of local contextual processing is necessary, especially in scenes where the style of the partially contaminated region is similar to that of the background. Thus, we retain the ACA module and use self-attention (K, Q, V components are all generated from the same input feature I) instead of cross-attention. The adaptive contextual features can be represented as:

$$\begin{aligned} &I_0, I_1, ..., I_{n-1} = \text{Split}(I), \\ &\hat{I}_0 = \text{DWConv}_{k\times k}(I_0), \\ &\hat{I}_i = \uparrow_{2^i} (\text{DWConv}_{k\times k}(\downarrow_{\frac{1}{2^i}} (I_i))), \\ &\hat{I} = \text{Concat}(\hat{I}_0, \hat{I}_1, ..., \hat{I}_{n-1}), \\ &\mathbb{G}\left[\hat{I}, I\right] = \phi(\hat{I}) \odot I. \end{aligned} \tag{9}$$

Fig. 4 shows feature representations at different scales and $\mathbb{G}[\hat{I}, I]$ aggregates rich contextual information. After obtaining $\hat{Q}$ through a convolution layer, the output $\hat{I}_a^G \in \mathbb{R}^{d \times c}$ of ACA can be formulated as:

$$\begin{aligned} &\hat{Q} = \text{Conv}_{1\times1}(\mathbb{G}\left[\hat{I}, I\right]), \\ &\hat{I}_a^G = \text{softmax}\left(\frac{\hat{Q}^T \cdot K}{\sqrt{d}}\right) \cdot V. \end{aligned} \tag{10}$$

D. High-frequency Omni-dimensional Dynamic Convolution

Due to the lack of mask guidance, blind image inpainting may struggle to detect contaminated regions that have semantic similarity to the background. Furthermore, current research [41] shows that the information lost in the process of downscaling is primarily high-frequency information. To better highlight contaminated regions and preserve texture, we propose a high-frequency omni-dimensional dynamic convolution (HODC) illustrated in Fig. 2(d) (the purple path), which utilizes edge features to amplify the representation of details.

Typically, dynamic convolution [42] selects n convolutional kernels W based on the input data, rather than using a single

(a) Input (b) VCNet [8] (c) TransHAE [9] (d) MAT [32] (e) OmniNet [3] (f) Ours(~~D~~) (g) Ours (h) GT

Fig. 6. Comparison with the state-of-the-art. These images come from CelebAMask-HQ [39], FFHQ [40] with various contamination patterns. Ours(~~D~~) refers to the configuration where the AFAN model is trained without employing the proposed mask region perception strategy denoted as D.

kernel in standard convolution. Later, the omni-dimensional dynamic convolution (ODC) [43] 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 ($\alpha_w \in \mathbb{R}$). Fig. 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 $I \in \mathbb{R}^{d\times c_{in}}$ from intermediate layers of AFAN, the ODC scheme can be formulated as:

$$\begin{aligned} \alpha_s, \alpha_c, \alpha_f, \alpha_w &= \mathbb{P}[I], \\ I_{odc} &= \sum_{i=1}^{n} (\alpha_{w_i} \odot \alpha_{f_i} \odot \alpha_{c_i} \odot \alpha_{s_i} \odot W_i) * I, \end{aligned} \tag{11}$$

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 with the Scharr kernels W_x and W_y, respectively:

$$W_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix}, \quad W_y = \begin{bmatrix} 3 & 10 & 3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix}. \tag{12}$$

Subsequently, the magnitude of the gradient feature S at each point is computed as follows:

$$S = \sqrt{(W_x * I)^2 + (W_y * I)^2}. \tag{13}$$

Fig. 5 shows that the edge features generated by Scharr filter can well represent the contours of the contaminated regions and the textures of the normal regions. To enhance the input features with the detected edge details, the weighted sum $I_s \in \mathbb{R}^{d\times c_{in}}$ of the original features and edge features can be formulated as:

$$I_s = \beta_1 I + \beta_2 S, \tag{14}$$

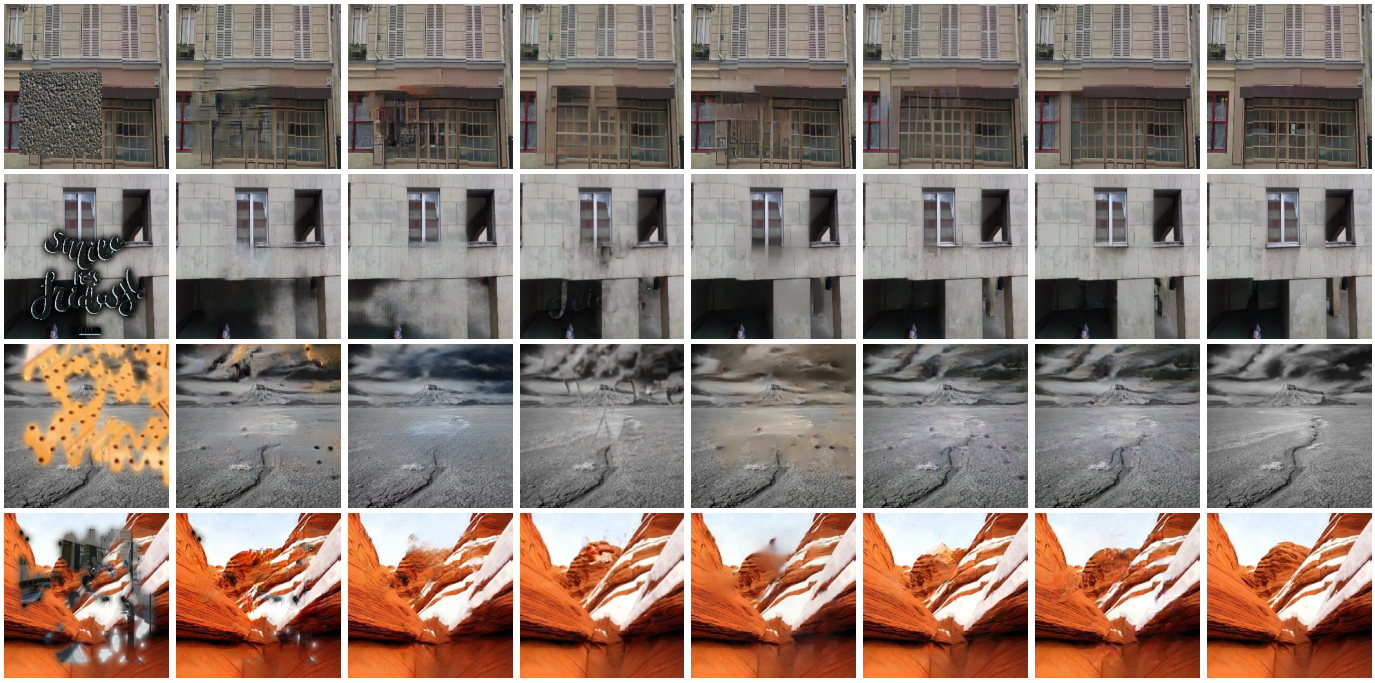
where β_1, β_2 are weights that control the contribution of the original image and the edge detail. In this work, we set $\beta_1 = 1$ and $\beta_2 = 0.5$, which means the enhanced image retains the original colors and brightness while emphasizing the texture. Finally, the output feature $I_{hodc} \in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

$$\begin{aligned} \hat{\alpha}_s, \hat{\alpha}_c, \hat{\alpha}_f, \hat{\alpha}_w &= \mathbb{P}[I_s], \\ I_{hodc} &= \sum_{i=1}^{n} (\hat{\alpha}_{w_i} \odot \hat{\alpha}_{f_i} \odot \hat{\alpha}_{c_i} \odot \hat{\alpha}_{s_i} \odot W_i) * I. \end{aligned} \tag{15}$$

Compared to methods that rely on additional high-frequency gating components to modulate the output features[44], [45], the HODC enhances the ability to sample fine texture details and identify the contaminated areas by adaptively selecting the omni-dimensional attention α for various convolutional kernels.

E. Loss Function

Taking into account the consistency between overall content and fine detail, AFAN applies four types of loss functions:



(a) Input (b) VCNet [8] (c) TransHAE [9] (d) MAT [32] (e) OmniNet [3] (f) Ours($\not{D}$) (g) Ours (h) GT

Fig. 7. Comparison with the state-of-the-art. These images come from Paris StreetView [46] and Places2 [47] with various contamination patterns. Ours($\not{D}$) refers to the configuration where the AFAN model is trained without employing the proposed mask region perception strategy denoted as D.

mean squared error (MSE) loss, perceptual loss, stochastic structural similarity (S3IM) loss [48], 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 function is as follows:

$$\mathcal{L}_{con} = \|I_o \odot I_{gt}\|_2^2, \tag{16}$$

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 [49].

$$\mathcal{L}_{perc} = \sum_i \|\Phi_i(I_o) - \Phi_i(I_{gt})\|_1, \tag{17}$$

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 majority of tasks involving image synthesis employ the Structural Similarity Index Measure (SSIM) loss, which captures local information from adjacent pixels using convolutional kernels. However, SSIM's ability to detect structural information in distant pixels is limited. To overcome this limitation, S3IM loss is a feasible scheme that randomly scrambles the pixel distribution of minibatch images to create non-local sets of pixels, and then SSIM is applied to these artificially constructed patches:

$$\mathcal{L}_{s3im} = 1 - \mathrm{S3IM}(I_o, I_{gt}). \tag{18}$$

In the training process of AFAN, the improved S3IM loss randomly scrambles the pixels within a single output image I_o (including the groundtruth) rather than using minibatch images in [48]. This innovation aims to enhance the detection of structural information across broader regions of each image, improving the quality and coherence of inpainting results.

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

$$\mathcal{L} = \min_G \max_D (\mathcal{L}_{con} + \lambda_1 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{s3im} + \lambda_3 \mathcal{L}_{adv} + \lambda_4 \mathcal{L}_s) \tag{19}$$

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

IV. Experiments

A. Implementation Details

The AFAN is evaluated using four public datasets including a range of subjects: CelebAMask-HQ [39] and FFHQ [40] for high-quality faces, Paris StreetView [46] and Places2 [47] 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 Fig. 8, we apply two contamination patterns: regular patterns and irregular patterns (including text-like patterns [50]), to simulate various types of blind images.

During the training phase, we use the Adam optimizer [51] 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 AFAN is developed using PyTorch and is trained on NVIDIA RTX 3090 GPUs.

B. Quantitative Evaluation

In the evaluation of inpainting results with various contamination patterns, AFAN is compared with state-of-the-art

TABLE I

QUANTITATIVE EVALUATIONS ON THE CELEBAMASK-HQ [39], FFHQ [40], PARIS STREETVIEW [46] AND PLACES2 [47] 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 AFAN IS TRAINED WITHOUT EMPLOYING THE PROPOSED MASK REGION PERCEPTION STRATEGY DENOTED AS D.

	Dataset	VCNet [8]	TransHAE [9]	MAT [32]	OmniNet [3]	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**
ℓ_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**

(a) (b) (c) (d) (e)

Fig. 8. A groundtruth image (a) can be subjected to contamination (b) using three distinct types of patterns: regular pattern (c), irregular pattern (d), and text-like pattern (e).

(a) Input (b) BF (c) b+HODC (d) c+ACA (e) GT

Fig. 9. Ablation study on different configurations of the AFAN for blind image inpainting. The experiment is conducted on the CelebAMask-HQ [39] dataset with regular contamination patterns.

such as VCNet [8], TransHAE [9], and OmniNet [3] for blind image inpainting. Meanwhile, a non-blind image inpainting method MAT [32] is applied as a comparative reference. These comparisons are conducted on testing datasets from CelebAMask-HQ [39], FFHQ [40], Places2 [47], and Paris StreetView [46]. 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) [52] and the Frechet Inception Score (FID) [53], are utilized to measure the perceptual quality of predicted images compared to the groundtruth images. As detailed in Table I, comparative experiments conducted on different datasets show that the proposed method outperforms existing approaches on most of the metrics.

C. Qualitative Evaluations

To validate the inpainting performance, Fig. 6 and Fig. 7 present a comparative analysis of the predicted results from different methods. As illustrated in Fig. 6, the inpainting result from VCNet seems to produce distorted structures, particularly noticeable around contaminated edge regions. TransHAE tends to produce 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, Fig. 7 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.

D. Ablation study

Fig. 6-7 (see columns f and g) and Table I (the last two columns) have shown that the proposed mask region percep-

TABLE II
ABLATION STUDY ON THE CELEBAMASK-HQ DATASET WITH REGULAR CONTAMINATION PATTERN.

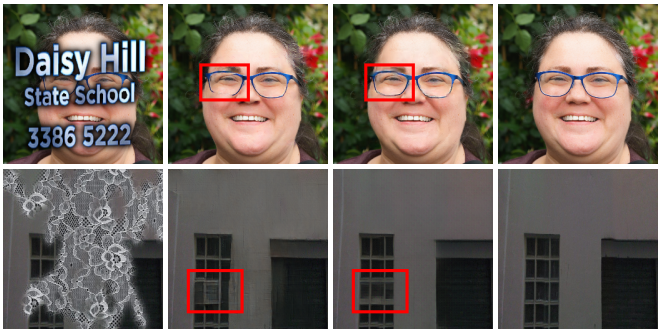
Methods	PSNR ↑	SSIM ↑	$\ell_1(\%)$ ↓	LPIPS ↓	FID ↓
BF	25.84	0.874	3.93	0.082	17.48
BF+HODC	26.13	0.891	3.47	0.079	15.27
BF+ACA	26.47	0.909	3.34	0.075	14.51
BF+HODC +ACA	**26.91**	**0.921**	**2.97**	**0.066**	**12.45**

TABLE III
ABLATION STUDY ON CELEBAMASK-HQ, FFHQ, PARIS STREETVIEW AND PLACES2 WITH VARIOUS CONTAMINATION PATTERNS AS INPUT. ~~ACA~~/~~HODC~~ REFERS TO THE CONFIGURATION WHERE THE AFAN MODEL IS TRAINED WITHOUT EMPLOYING THE ACA/HODC MODULE.

	Model	FFHQ	Paris StreetView
PSNR ↑	AFAN(~~ACA~~)	26.5408	25.3473
	AFAN	27.1040	26.9927
SSIM ↑	AFAN(~~ACA~~)	0.9033	0.8613
	AFAN	0.9124	0.8724
$\ell_1(\%)$ ↓	AFAN(~~ACA~~)	3.7346	3.8723
	AFAN	2.1642	2.8544
LPIPS ↓	AFAN(~~ACA~~)	0.0760	0.0924
	AFAN	0.0459	0.0805
FID ↓	AFAN(~~ACA~~)	12.8712	39.3578
	AFAN	10.3784	34.9745
	Model	Places2	CelebAMask-HQ
PSNR ↑	AFAN(~~HODC~~)	26.5374	27.9465
	AFAN	26.7409	28.2603
SSIM ↑	AFAN(~~HODC~~)	0.8716	0.9201
	AFAN	0.8983	0.9387
$\ell_1(\%)$ ↓	AFAN(~~HODC~~)	2.2062	1.9305
	AFAN	2.1702	1.8316
LPIPS ↓	AFAN(~~HODC~~)	0.0779	0.0457
	AFAN	0.0722	0.0411
FID ↓	AFAN(~~HODC~~)	22.0278	9.2674
	AFAN	20.5134	8.4829

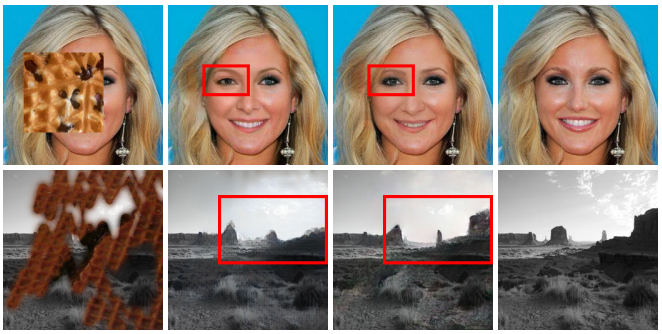
tion strategy D significantly reduces the presence of contaminant artifacts. In this subsection, we continue to analyze how the other proposed modules (ACA block, HODC) contribute to the final performance of image inpainting. Specifically, we verify the effectiveness of the AFAN backbone framework (BF) by removing HODC and replacing ACA blocks of the generator with standard transformer blocks. Following this, the HODC layers and ACA operation are integrated into the backbone in turn, and this operation allows us to assess the contributions of these components to the overall performance. Fig. 9 shows that these components sequentially 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 II illustrates that our proposed modules demonstrably enhance the performance in the task of blind image inpainting.

To further assess ACA's contribution to improve local content representation, ablation studies are carried out using both full AFAN and ACA-free AFAN. Fig. 10 and Table III illustrate the results for different datasets. It is evident



(a) Input (b) ~~ACA~~ (c) AFAN (d) GT

Fig. 10. Ablation study of the ACA strategy. The experiment is conducted on FFHQ [40] and Paris StreetView [46]. ~~ACA~~ refers to the configuration where the AFAN model is trained without employing the ACA module.



(a) Input (b) ~~HODC~~ (c) AFAN (d) GT

Fig. 11. Ablation study of the HODC strategy. The experiment is conducted on CelebAMask-HQ [39] and Places2 [47]. ~~HODC~~ refers to the configuration where the AFAN model is trained without HODC layers.

that ACA plays a crucial role in improving the precision of local feature identification while concurrently preserving rich levels of detail. Similarly, the HODC employs edge features calculated by the Scharr operator to enhance the expression of details, and its efficacy is further confirmed by the visual results presented in Fig. 11.

V. CONCLUSION

This paper presents AFAN, 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, AFAN introduces adaptive 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 AFAN achieves superior results in blind image inpainting for various contamination. The proposed AFAN 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] N. Cai, Z. Su, Z. Lin, H. Wang, Z. Yang, and B. W.-K. Ling, "Blind inpainting using the fully convolutional neural network," The Visual Computer, vol. 33, pp. 249–261, 2017. 1, 2

[2] S. Zhang, R. He, Z. Sun, and T. Tan, "Demeshnet: Blind face inpainting for deep meshface verification," IEEE TIFS, vol. 13, no. 3, pp. 637–647, 2017. 1, 2

[3] S. S. Phutke, A. Kulkarni, S. K. Vipparthi, and S. Murala, "Blind image inpainting via omni-dimensional gated attention and wavelet queries," in CVPR, 2023, pp. 1251–1260. 1, 2, 6, 7, 8

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in NeurIPS. USA: Curran Associates, 2014. 1

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in NeurIPS, vol. 30. USA: Curran Associates, 2017. 1

[6] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in CVPR, 2022, pp. 5728–5739. 1, 2

[7] Y. Deng, S. Hui, S. Zhou, D. Meng, and J. Wang, "T-former: An efficient transformer for image inpainting," in ACM MM, 2022, pp. 6559–6568. 1

[8] H. Zhao, Z. Gu, B. Zheng, and H. Zheng, "Transcnn-hae: Transformer-cnn hybrid autoencoder for blind image inpainting," in ACM MM, 2022, pp. 6813–6821. 1, 2, 6, 7, 8

[9] Y. Wang, Y.-C. Chen, X. Tao, and J. Jia, "Vcnet: A robust approach to blind image inpainting," in ECCV. Springer, 2020, pp. 752–768. 1, 2, 6, 7, 8

[10] Y. Liu, J. Pan, and Z. Su, "Deep blind image inpainting," in IScIDE. Springer, 2019, pp. 128–141. 1, 2

[11] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao, and F. Wen, "Bringing old photos back to life," in CVPR, 2020, pp. 2747–2757. 1

[12] J. Wang, S. Chen, Z. Wu, and Y.-G. Jiang, "Ft-tdr: Frequency-guided transformer and top-down refinement network for blind face inpainting," IEEE TMM, 2022. 1, 2

[13] J. Wang, C. Yuan, B. Li, Y. Deng, W. Hu, and S. Maybank, "Self-prior guided pixel adversarial networks for blind image inpainting," IEEE PAMI, 2023. 1, 2

[14] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in Siggraph. USA: ACM, 2000, pp. 417–424. 2

[15] T. F. Chan and J. Shen, "Nontexture inpainting by curvature-driven diffusions," JVCIR, vol. 12, no. 4, pp. 436–449, 2001. 2

[16] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patchmatch: A randomized correspondence algorithm for structural image editing," ToG, vol. 28, no. 3, p. 24, 2009. 2

[17] S. Darabi, E. Shechtman, C. Barnes, D. B. Goldman, and P. Sen, "Image melding: Combining inconsistent images using patch-based synthesis," ToG, vol. 31, no. 4, pp. 1–10, 2012. 2

[18] R. Gao and K. Grauman, "On-demand learning for deep image restoration," in ICCV. Italy: IEEE Computer Society, 2017, pp. 1086–1095. 2

[19] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," ToG, vol. 36, no. 4, pp. 1–14, 2017. 2

[20] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in ECCV. USA: Springer, 2018, pp. 85–100. 2

[21] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Structure guided image inpainting using edge prediction," in ICCV Workshops. Korea: IEEE Computer Society, 2019, pp. 0–0. 2

[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv preprint arXiv:1511.06434, 2015. 2

[23] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in ICCV. Korea: IEEE Computer Society, 2019, pp. 4471–4480. 2

[24] J. Liu, S. Yang, Y. Fang, and Z. Guo, "Structure-guided image inpainting using homography transformation," IEEE TMM, vol. 20, no. 12, pp. 3252–3265, 2018. 2

[25] R. Zhang, W. Quan, Y. Zhang, J. Wang, and D.-M. Yan, "W-net: Structure and texture interaction for image inpainting," IEEE TMM, vol. 25, pp. 7299–7310, 2022. 2

[26] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in CVPR. USA: IEEE Computer Society, 2018, pp. 5505–5514. 2

[27] N. Wang, J. Li, L. Zhang, and B. Du, "Musical: Multi-scale image contextual attention learning for inpainting." in IJCAI. China: AAAI Press, 2019, pp. 3748–3754. 2

[28] H. Wu, J. Zhou, and Y. Li, "Deep generative model for image inpainting with local binary pattern learning and spatial attention," IEEE TMM, vol. 24, pp. 4016–4027, 2021. 2

[29] Y. Yu, F. Zhan, R. Wu, J. Pan, K. Cui, S. Lu, F. Ma, X. Xie, and C. Miao, "Diverse image inpainting with bidirectional and autoregressive transformers," in ACM MM, 2021, pp. 69–78. 2

[30] C. Zheng, T.-J. Cham, J. Cai, and D. Phung, "Bridging global context interactions for high-fidelity image completion," in CVPR. USA: IEEE Computer Society, 2022, pp. 11 512–11 522. 2

[31] Q. Liu, Z. Tan, D. Chen, Q. Chu, X. Dai, Y. Chen, M. Liu, L. Yuan, and N. Yu, "Reduce information loss in transformers for pluralistic image inpainting," in CVPR. USA: IEEE Computer Society, 2022, pp. 11 347–11 357. 2

[32] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in CVPR, 2022, pp. 10 758–10 768. 2, 6, 7, 8

[33] S. Chen, A. Atapour-Abarghouei, and H. P. Shum, "Hint: High-quality inpainting transformer with mask-aware encoding and enhanced attention," IEEE TMM, 2024. 2

[34] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," IEEE TIFS, vol. 15, pp. 144–159, 2019. 4

[35] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in CVPR, 2023, pp. 20 606–20 615. 4

[36] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," IEEE TIFS, vol. 7, no. 3, pp. 868–882, 2012. 4

[37] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in CVPR. USA: IEEE Computer Society, 2017, pp. 1125–1134. 4

[38] H. Lin, X. Cheng, X. Wu, and D. Shen, "Cat: Cross attention in vision transformer," in ICME. IEEE, 2022, pp. 1–6. 5

[39] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in CVPR. USA: IEEE Computer Society, 2020, pp. 5549–5558. 6, 7, 8, 9

[40] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in CVPR, 2019, pp. 4401–4410. 6, 7, 8, 9

[41] M. Xiao, S. Zheng, C. Liu, Y. Wang, D. He, G. Ke, J. Bian, Z. Lin, and T.-Y. Liu, "Invertible image rescaling," in ECCV. UK: Springer, 2020, pp. 126–144. 5

[42] Y. Chen, X. Dai, M. Liu, D. Chen, L. Yuan, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in CVPR, 2020, pp. 11 030–11 039. 5

[43] C. Li, A. Zhou, and A. Yao, "Omni-dimensional dynamic convolution," arXiv preprint arXiv:2209.07947, 2022. 6

[44] Y. Yu, F. Zhan, S. Lu, J. Pan, F. Ma, X. Xie, and C. Miao, "Wavefill: A wavelet-based generation network for image inpainting," in ICCV. Canada: IEEE Computer Society, 2021, pp. 14 114–14 123. 6

[45] B. Li, B. Zheng, H. Li, and Y. Li, "Detail-enhanced image inpainting based on discrete wavelet transforms," Signal Processing, vol. 189, p. 108278, 2021. 6

[46] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in CVPR. USA: IEEE Computer Society, 2016, pp. 2536–2544. 7, 8, 9

[47] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," TPAMI, vol. 40, no. 6, pp. 1452–1464, 2017. 7, 8, 9

[48] Z. Xie, X. Yang, Y. Yang, Q. Sun, Y. Jiang, H. Wang, Y. Cai, and M. Sun, "S3im: Stochastic structural similarity and its unreasonable effectiveness for neural fields," in ICCV, 2023, pp. 18 024–18 034. 7

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014. 7

[50] X. Xu, Z. Zhang, Z. Wang, B. Price, Z. Wang, and H. Shi, "Rethinking text segmentation: A novel dataset and a text-specific refinement approach," in CVPR, 2021, pp. 12 045–12 055. 7

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014. 7

[52] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in CVPR, 2018, pp. 586–595. 8

[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," NeurIPS, vol. 30, 2017. 8