IEEE Transactions on Multimedia

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

Journal:	*IEEE Transactions on Multimedia*
Manuscript ID	MM-020898.R1
Manuscript Type:	Regular Paper (S1)
Date Submitted by the Author:	23-Nov-2024
Complete List of Authors:	Wang, Jiahao; Tianjin University, College of Intelligence and Computing Pan, Gang; Tianjin University, College of Intelligence and Computing & School of New Media and Communication Sun, Di; Tianjin University of Science and Technology, College of Artificial Intelligence Li, Jinyuan; Tianjin University, School of New Media and Communication Zhang, Jia-Wan; Tianjin University, School of Computer Software
Subject Category Please select at least one subject category that best reflects the scope of your manuscript:	SIGNAL PROCESSING FOR MULTIMEDIA APPLICATIONS, EMERGING TOPICS IN MULTIMEDIA
EDICS:	1-IVGAS Image/Video/Graphics Analysis and Synthesis < 1 SIGNAL PROCESSING FOR MULTIMEDIA APPLICATIONS, 9-DLMP Deep Learning for Multimedia Processing < 9 EMERGING TOPICS IN MULTIMEDIA

Author's Responses to Custom Submission Questions

Is this manuscript a resubmission of, or related to, a previously rejected manuscript?	No
If "Yes", specify the publication venue and manuscript ID of the previous submission and upload a supporting document detailing how the resubmission has addressed the concerns raised during the previous review. If this does not apply, type N/A.	N/A
Is this manuscript an extended version of a conference publication?	No
If "Yes", provide the full citation of the conference submission or publication. If this does not apply, type N/A.	N/A
Is this manuscript related to any other papers of the authors that are either published, accepted for publication, or currently under review, and that are not included among the references cited in the manuscript?	No
If "Yes", please list these papers below. Except for permitted preprints, explain why these papers are not included among the references cited in the manuscript and how they are different from the manuscript. Include any unpublished papers as "Supporting Documents". If this does not apply, type N/A.	N/A
What is the contribution of this paper, within the scope of Transactions on Multimedia?	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 (AFAN), which can perform inpainting operations in an end-to-end manner, utilizing the proposed mask region perception strategy. We design an adaptive contextual attention (ACA) 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 (HODC), which incorporates edge features to improve the representation of details.
Why is the contribution significant (What impact will it have)?	The contribution is significant because it introduces innovative approaches that address key challenges in blind image inpainting, leading to advancements with

	broad implications for various applications. 1) Improved Image Synthesis Accuracy: By integrating adversarial training with forgery region detection, the model can better identify and focus on contaminated areas, resulting in more accurate and realistic inpainting. This advancement is crucial for tasks requiring high-quality image restoration, such as old photos restoration and historical artifact preservation. 2) Enhanced Reconstruction Capabilities: The adaptive contextual attention algorithm effectively captures both long-range dependencies and local contextual features. This dual-focus mechanism improves the model's ability to handle complex image structures, making it more versatile for diverse scenarios. 3) Refined Detail Representation: The high-frequency omni-dimensional dynamic convolution leverages edge features to enhance the representation of fine details. This improvement addresses the challenge of preserving high-frequency information, which is vital for producing visually appealing and precise outputs. These contributions collectively push the boundaries of image inpainting, offering solutions that are more accurate, adaptable, and detail-oriented, with potential impacts in fields like computer vision, artificial intelligence, and multimedia processing.
What are the three papers in the published literature most closely related to this paper? Please provide full citation details, including DOI references where possible.	1. 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 2. 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 3. H. Zhao, Z. Gu, B. Zheng, and H. Zheng, "Transcnn-hae: Transformercnn hybrid autoencoder for blind image inpainting," in ACM MM, 2022, pp. 6813-6821. 1, 2, 6, 7, 8
What is distinctive/new about the current paper relative to these previously published works?	1. Novel Mask Region Perception Strategy. Unlike existing works that may focus solely on attention mechanisms or hybrid architectures, this paper introduces a unique mask region perception strategy. By combining adversarial training with forgery detection, the method enhances the model's ability to accurately localize and perceive corrupted areas, which is a

	significant improvement in blind image inpainting. 2.Attention-Driven Forgery Adversarial Network (AFAN). The proposed AFAN introduces adaptive contextual attention (ACA) blocks to achieve feature modulation effectively. While previous works, such as omni-dimensional gated attention or transformer-CNN hybrids, explore attention mechanisms, this paper uniquely focuses on adaptively capturing both long-range and local dependencies to enhance reconstruction performance. 3.High-Frequency Omni-Dimensional Dynamic Convolution (HODC). This paper uniquely integrates edge feature enhancement through HODC to improve the representation of high-frequency details. In contrast, prior methods like wavelet queries or general convolutional approaches do not explicitly prioritize high-frequency and edge information, making this a novel contribution to detail preservation and fidelity. 4.End-to-End Inpainting with Comprehensive Design. While studies such as VcNet and TransCNN-HAE contribute robust hybrid frameworks, this paper offers a more cohesive end-to-end approach with attention, adversarial learning, and convolution integrated seamlessly. The holistic design improves performance in diverse and challenging scenarios. The combination of the mask region perception strategy, the novel ACA blocks, and the high-frequency detail-enhancing HODC sets this paper apart by tackling blind image inpainting with a more targeted and comprehensive methodology, delivering improvements in both perceptual quality and reconstruction accuracy.

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 aimed at reconstructing corrupted regions without relying on mask information. Due to the lack of mask priors, previous methods usually integrate a mask prediction network in the initial phase, followed by an inpainting backbone. However, this multi-stage generation process may result in feature misalignment. While recent end-to-end generative methods bypass the mask prediction step, they typically struggle with weak perception of contaminated regions and introduce structural distortions. This study presents a novel mask region perception strategy for blind image inpainting by combining adversarial training with forgery detection. To implement this strategy, we propose an attention-driven forgery adversarial network (AFAN), which leverages adaptive contextual attention (ACA) blocks for effective feature modulation. Specifically, within the generator, ACA employs self-attention to enhance content reconstruction by utilizing the rich contextual information of adjacent tokens. In the discriminator, ACA utilizes cross-attention with noise priors to guide adversarial learning for forgery detection. Moreover, we design a high-frequency omni-dimensional dynamic convolution (HODC) based on edge feature enhancement to improve detail representation. Extensive evaluations across multiple datasets demonstrate that the proposed AFAN model outperforms existing generative methods in blind image inpainting, particularly in terms of quality and texture fidelity.**

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

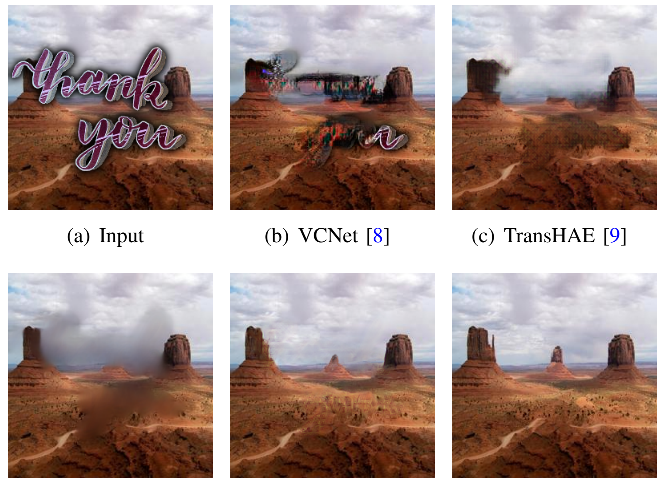
I. Introduction

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]–[3] usually employ general inpainting frameworks and combine with Generative Adversarial Networks (GANs) [4], transformer blocks [5]–[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).

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).



(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 [8]–[13] decompose blind image inpainting into two sub-tasks: mask prediction and universal image inpainting. Previous works [8], [10], [11] 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 [9] 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 the 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 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 capable of performing inpainting operations in an end-to-end manner, leveraging 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-based [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 struggle to handle distortions involving extensive or complex content. With the advent of deep learning, it has become the dominant technique in the field of image inpainting. Related works [18]–[21] commonly utilize the encoder-decoder architectures and enhance contextual understanding through advanced modules, such as GAN loss [22], gated convolution [23], contextual attention mechanisms [24], [25]. Although effective in addressing abnormal features, these methods face challenges in reconstructing large missing regions. To capture information located far apart spatially, mainstream methods [26]–[28] integrate pixel-wise attention blocks into the models, primarily reinforcing global context. Recently, the focus has shifted towards transformer-based methods [6], [29]–[33], which are suitable for non-local modeling and are highly effective at understanding and reconstructing image content across large spatial extents. Despite their strengths, these methods typically rely on mask information for inpainting, limiting their applicability in scenarios where such mask data is unavailable. Consequently, some researchers [34]–[36] have explored the use of text features as an alternative to mask information for image and video frame inpainting. In response to these challenges, a new approach known as blind image inpainting has emerged, enabling the recovery of corrupted regions without requiring 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.* [8] 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 [9] merges global modeling of the transformer and local modeling of CNN into a single

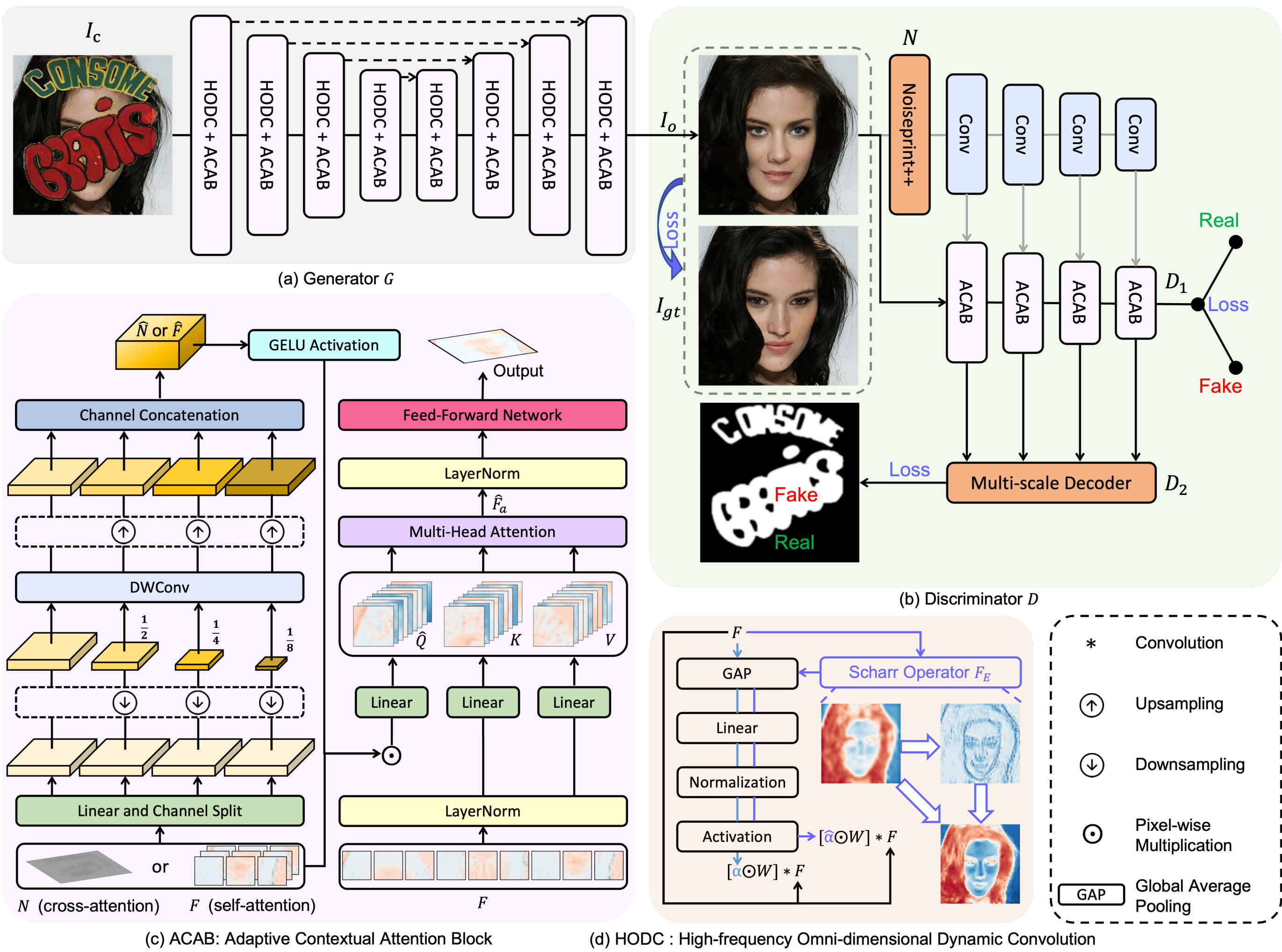


Fig. 2. Framework Overview. The AFAN consists of a generator G and a two-branch discriminator D. G integrates adaptive contextual attention (ACA) blocks to capture both long-range dependencies and local contextual features effectively, and the high-frequency omni-dimensional dynamic convolution (HODC)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

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 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.

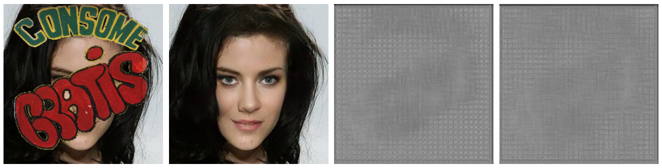
III. Approach

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:

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

where $\odot$ is pixel-wise multiplication and $S \in \mathbb{R}^{h\times w\times 3}$ is a visual signal (e.g., constant values, random noise or graffiti). $\mathbb{G}[\cdot]$ 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



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

Fig. 3. Noise-sensitive 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.

renders the contaminated areas less noticeable. The following sections will describe the framework architecture and image computation.

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 [37], Noiseprint++ [38], and SRM filtering [39]. 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 fed into the discriminator D as input. The encoder of D consists of downsampling layers and ACA blocks. To enhance the robustness of forgery detection, the noise-sensitive fingerprints N are integrated into the image features through the cross-attention mechanism of the ACA blocks. As indicated in Fig. 4, the integration of noise fingerprints N significantly enhances the discriminator's capability to identify forged regions. Subsequently, 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 [40] to optimize both

(a) Input (b) Output (c) SA(I_o) (d) CA(I_o, N)

Fig. 4. Forgery discrimination heatmaps generated using self-attention SA and cross-attention CA of the ACA block. (b) shows the image recovered by AFAN, and (d) indicates the fusion of noise fingerprints N and inpainted images I_o via cross-attention.

the projected discriminator D and the generator G. Thus the objective function for the GAN process is expressed as:

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

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

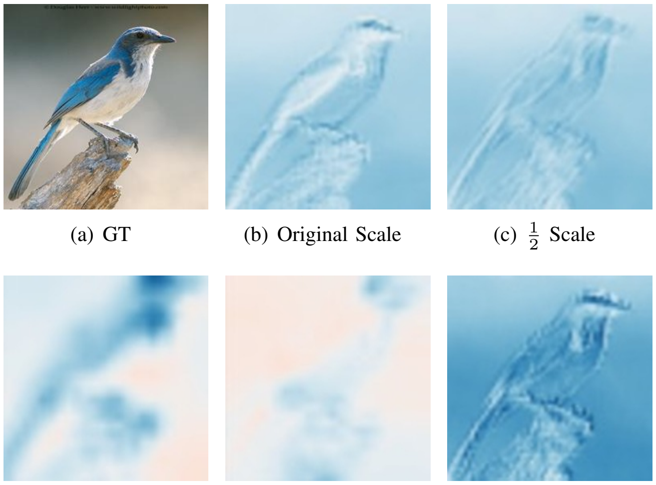
$$\begin{aligned}\mathcal{L}^D_{forg} &= \mathbb{E}_{I_{gt}}[ReLU(1 - D_2(I_{gt}, N))] \\ &\quad + \mathbb{E}_{I_o}[ReLU(1 - D_2(I_o, N) \odot (1 - M)] \\ &\quad + \mathbb{E}_{I_o}[ReLU(1 + D_2(I_o, N) \odot M)], \\ \mathcal{L}^G_{forg} &= -\mathbb{E}_{I_o}[D_2(I_o, N) \odot M].\end{aligned} \tag{3}$$

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 $F \in \mathbb{R}^{d \times c}$ (d is spatial size and c is channel) from intermediate layers of AFAN, the attention first converts F into queries Q,



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

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

Fig. 5. Multi-scale feature representation. (b)-(e) are the contextual feature representations sampled at different spatial scales in the ACA. (f) represents the gated mechanism $\mathcal{G}[\hat{F}, F]$ that adaptively fuse these multi-scale features.

keys K, and values V using respective linear matrices , and the output $F_a \in \mathbb{R}^{d \times c}$ is formulated as follows:

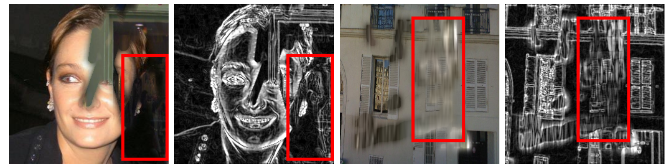
$$F_a = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d}}\right) \cdot V. \tag{4}$$

Building on this, CAT [41] 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 Q_N input and the inpainted image I_o as a key K and value V 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 ($d \times d$):

$$F_a^D = V \cdot \text{softmax}\left(\frac{Q_N^T \cdot K}{\sqrt{d}}\right). \tag{5}$$

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. 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 features $N \in \mathbb{R}^{d \times c}$ extracted from noise fingerprints 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 (DW-Conv) with kernel size $k = 2n - 1$ to collect local contextual information, while the rest parts N_i ($i \in [1, n-1]$) are down-



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

Fig. 6. 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.

sampled 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 $\mathcal{G}[\cdot]$ to adaptively fuse them:

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

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 $\hat{F}_a^D \in \mathbb{R}^{d \times c}$ of ACA is calculated as follows:

$$\begin{aligned} \hat{Q}_N &= \text{Conv}_{1 \times 1}(\mathcal{G}\left[\hat{N}, N\right]), \\ \hat{F}_a^D &= V \cdot \text{softmax}\left(\frac{\hat{Q}_N^T \cdot K}{\sqrt{d}}\right). \end{aligned} \tag{8}$$

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 F via linear layers) instead of cross-attention. The adaptive contextual features can be represented as:

$$\begin{aligned} F_0, F_1, ..., F_{n-1} &= \text{Split}(F), \\ \hat{F}_0 &= \text{DWConv}_{k \times k}(F_0), \\ \hat{F}_i &= \uparrow_{2^i} (\text{DWConv}_{k \times k}(\downarrow_{\frac{1}{2^i}} (F_i))), \\ \hat{F} &= \text{Concat}(\hat{F}_0, \hat{F}_1, ..., \hat{F}_{n-1}), \\ \mathcal{G}\left[\hat{F}, F\right] &= \phi(\hat{F}) \odot F. \end{aligned} \tag{9}$$

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

Fig. 7. Comparison with the state-of-the-art. These images come from CelebAMask-HQ [42], FFHQ [43] with various contamination patterns.

Fig. 5 shows feature representations at different scales and $\mathcal{G}[\hat{F}, F]$ aggregates rich contextual information. After obtaining $\hat{Q}$ through a convolution layer, the output $\hat{F}_a^G \in \mathbb{R}^{d \times c}$ of ACA can be formulated as:

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

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 [44] 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. For instance, Fig. 6 indicates that the edge features can well represent the contours of the contaminated regions and the textures of the normal regions in the input image I_c.

Typically, dynamic convolution [45] selects n convolutional kernels W based on the input data, rather than using a single kernel in standard convolution. Later, the omni-dimensional dynamic convolution (ODC) [46] 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 $F \in \mathbb{R}^{d\times c_{in}}$ from intermediate layers of AFAN, the ODC scheme can be formulated as:

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

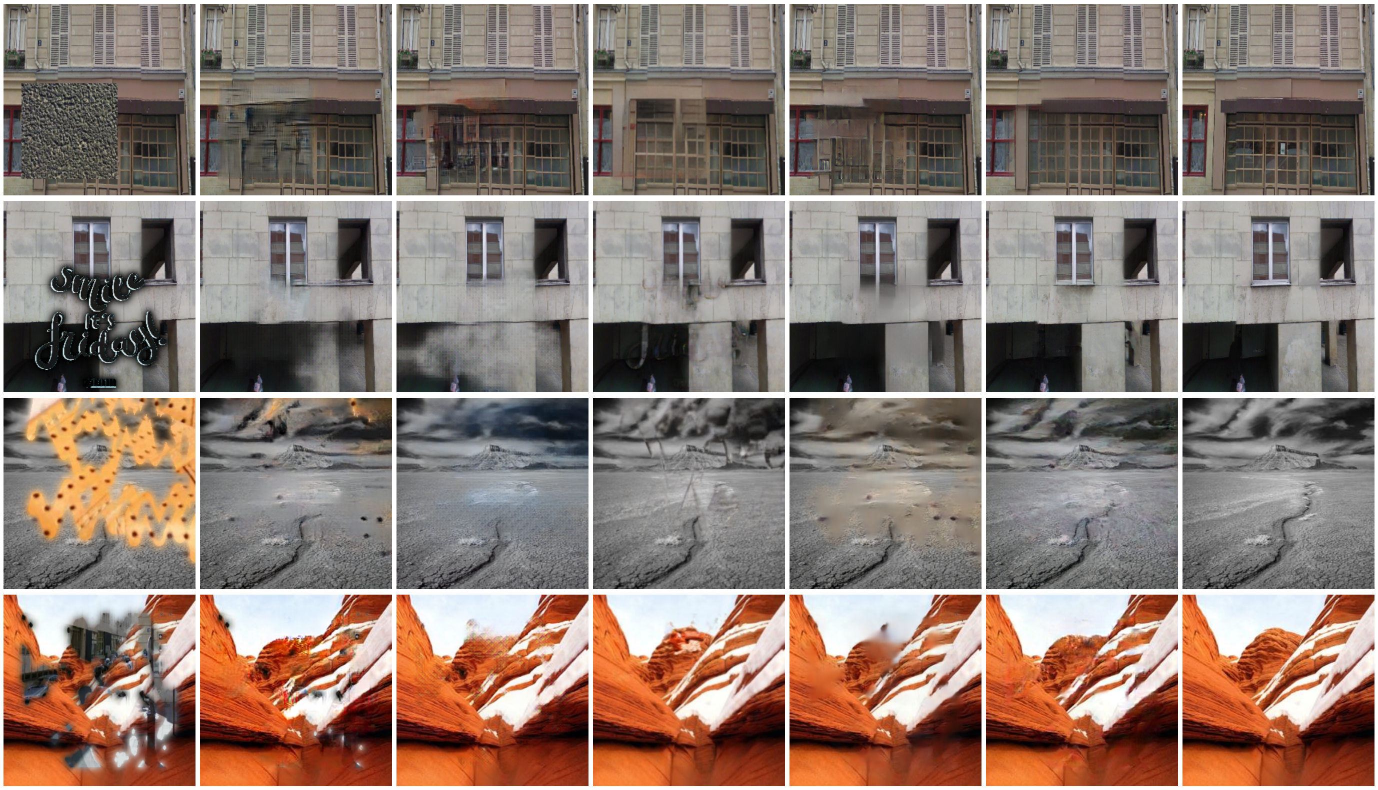
where $F_{odc} \in \mathbb{R}^{d\times c_{out}}$ is the output features, $*$ is the convolution operation.

To amplify the representation of details, HODC employs images created through edge detection (e.g., Scharr filter [47]) to augment the fine details in the input features. Specifically, the Scharr operator computes the gradients of F 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 E at each point is computed as follows:

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



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

Fig. 8. Comparison with the state-of-the-art. These images come from Paris StreetView [48] and Places2 [49] with various contamination patterns.

To enhance the input features with the detected edge details, the weighted sum $F_E \in \mathbb{R}^{d \times c_{in}}$ of the original features and edge features can be formulated as:

$$F_E = \beta_1 F + \beta_2 E, \tag{14}$$

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 $F_{hodc} \in \mathbb{R}^{d \times c_{out}}$ can be represented as:

$$\begin{gathered} \hat{\alpha}_s, \hat{\alpha}_c, \hat{\alpha}_f, \hat{\alpha}_w = \mathbb{P}\left[F_E\right], \\ F_{hodc} = \sum_{i=1}^{n} (\hat{\alpha}_{w_i} \odot \hat{\alpha}_{f_i} \odot \hat{\alpha}_{c_i} \odot \hat{\alpha}_{s_i} \odot W_i) * F. \end{gathered} \tag{15}$$

Fig. 10 visualizes the feature maps generated by each component using ODC and HODC, respectively. The HODC module incorporates edge features to strengthen the encoder's capability in identifying contaminated areas while enhancing the decoder's proficiency in capturing fine texture details.

E. Loss Function

Taking into account the consistency between overall content and fine detail, AFAN applies four types of loss functions: mean squared error (MSE) loss, perceptual loss, stochastic structural similarity (S3IM) loss [50], 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:

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

Fig. 9. 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).

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

where $\|\cdot\|_2$ 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 [51].

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

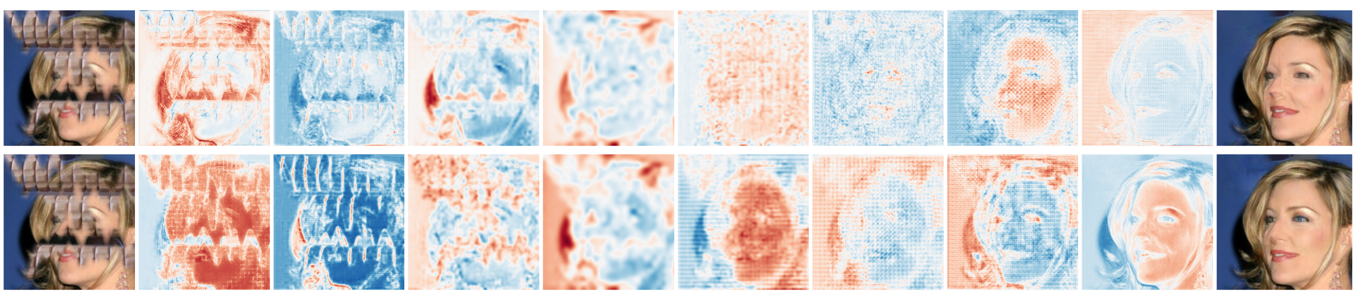
where Φ_i represents the output feature map of the i-th layer in VGG-16, corresponding to the activation layers: $ReLU1\_1$, $ReLU2\_1$, $ReLU3\_1$, $ReLU4\_1$, and $ReLU5\_1$.

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

TABLE I

QUANTITATIVE EVALUATIONS ON THE CELEBAMASK-HQ [42], FFHQ [43], PARIS STREETVIEW [48] AND PLACES2 [49] WITH VARIOUS CONTAMINATION PATTERNS AS INPUT. ↓ INDICATES THE LOWER THE BETTER WHILE ↑ MEANS THE HIGHER THE BETTER.

	Dataset	VCNet [9]	TransHAE [8]	MAT [32]	OmniNet [3]	Ours
PSNR ↑	CelebAMask-HQ	24.4288	27.3579	26.5847	24.8500	**28.2603**
	FFHQ	23.2432	26.9964	25.7812	23.1101	**27.1040**
	Paris StreetView	23.7850	24.9231	25.0484	22.8219	**26.9927**
	Places2	25.0681	25.4577	26.0403	24.8325	**26.7409**
SSIM ↑	CelebAMask-HQ	0.8871	0.9005	0.9157	0.8997	**0.9387**
	FFHQ	0.8988	**0.9163**	0.9112	0.9010	0.9124
	Paris StreetView	0.8275	0.8626	0.8713	0.8025	**0.8724**
	Places2	0.8615	0.8882	0.8741	0.8291	**0.8983**
$\ell_1(\%)$ ↓	CelebAMask-HQ	4.3712	2.6468	3.8901	4.9374	**1.8316**
	FFHQ	4.2832	**2.0420**	3.7285	5.0538	2.1642
	Paris StreetView	5.8475	3.1092	3.8565	4.4269	**2.8544**
	Places2	4.7277	3.0596	2.3656	3.8230	**2.1702**
LPIPS ↓	CelebAMask-HQ	0.1380	0.0722	0.0651	0.1424	**0.0411**
	FFHQ	0.1125	0.0866	0.0874	0.1840	**0.0459**
	Paris StreetView	0.1653	0.0991	**0.0795**	0.2173	0.0805
	Places2	0.0921	0.0941	0.0825	0.1480	**0.0722**
FID ↓	CelebAMask-HQ	13.2764	11.9616	10.9558	14.9926	**8.4829**
	FFHQ	13.3812	11.6033	12.0317	15.2021	**10.3784**
	Paris StreetView	52.1438	35.8904	38.3674	43.4504	**34.9745**
	Places2	27.2467	23.4471	24.4273	23.1912	**20.5134**



(a) Input (b) C_1 (c) C_2 (d) C_3 (e) C_4 (f) C_5 (g) C_6 (h) C_7 (i) C_8 (j) Output

Fig. 10. Feature map visualization of the generator. C represents the 8 transformer-style components. Row 1 shows outputs employing the ODC module, while Row 2 shows outputs employing the HODC module. $C_1 - C_4$ are encoder components and C_5–C_8 are decoder components.

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 [50]. 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:

$$\mathcal{L} = \mathcal{L}_{con} + \lambda_1 \mathcal{L}_{perc} + \lambda_2 \mathcal{L}_{s3im} + \lambda_3 \mathcal{L}_{adv} + \lambda_4 \mathcal{L}_{forg} \tag{19}$$

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

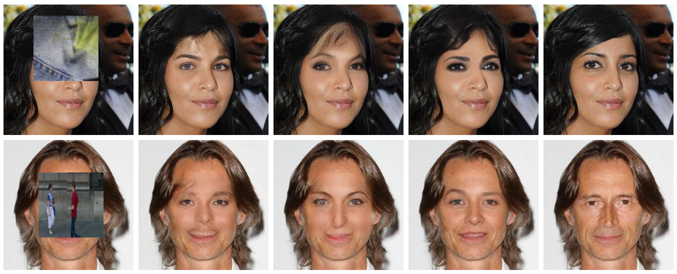
A. Implementation Details

The AFAN is evaluated using four public datasets including a range of subjects: CelebAMask-HQ [42] and FFHQ [43] for high-quality faces, Paris StreetView [48] and Places2 [49] 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. 9, we apply two contamination patterns: regular patterns and irregular patterns (including text-like patterns [52]), to simulate various types of blind images.

During the training phase, we use the Adam optimizer [53] 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



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

Fig. 11. Ablation study on different configurations of the AFAN for blind image inpainting. The experiment is conducted on the CelebAMask-HQ [42] dataset with regular contamination patterns.

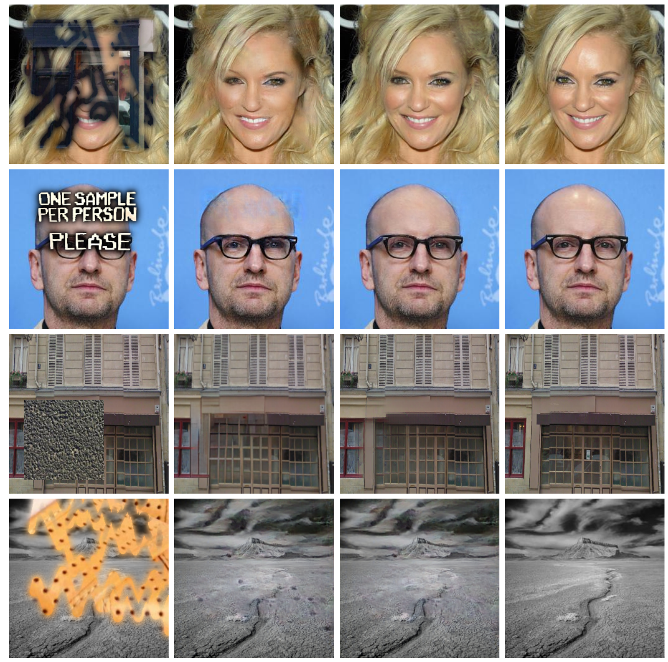
TABLE II
ABLATION STUDY ON THE CELEBAMASK-HQ [42] DATASET WITH REGULAR CONTAMINATION PATTERN.

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**

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 [42], FFHQ [43], Places2 [49], and Paris StreetView [48]. 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) [54] and the Frechet Inception Score (FID) [55], 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. 7 and Fig. 8 present a comparative analysis of the predicted results from different methods. As illustrated in Fig. 7, 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. 8 shows similar results on the



(a) Input (b) w/o D (c) AFAN (d) GT

Fig. 12. Ablation study of the discriminator D.The experiment is conducted on four datasets with contamination. w/o D refers to the configuration in which the AFAN model is trained without employing the proposed mask region perception strategy denoted as D.

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.

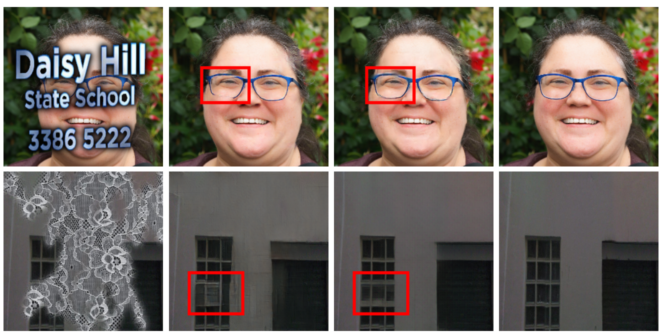
D. Ablation study

In this subsection, we analyze how the proposed modules (ACA block, HODC) contribute to the final performance of image inpainting. Specifically, we evaluate the effectiveness of the AFAN backbone framework (BF) by removing the HODC module and replacing the ACA blocks in the generator with standard transformer blocks. Following this, the HODC layers and ACA scheme are progressively integrated into the backbone, enabling us to assess their individual contributions to the overall performance systematically. As shown in Fig. 11, these components sequentially enhance the generation of reasonable contextual content and fine texture details on the CelebAMask-HQ [42] dataset. Note that this dataset adopts regular contamination patterns, which are referred 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 analyze the contribution of each module to the overall performance, we train a series of variant AFANs: i) without (denoted as w/o) the proposed mask region perception strategy, which is enabled by the discriminator D; ii) without employing the ACA scheme; iii) without incorporating the HODC layers. Quantitative comparisons between these AFAN

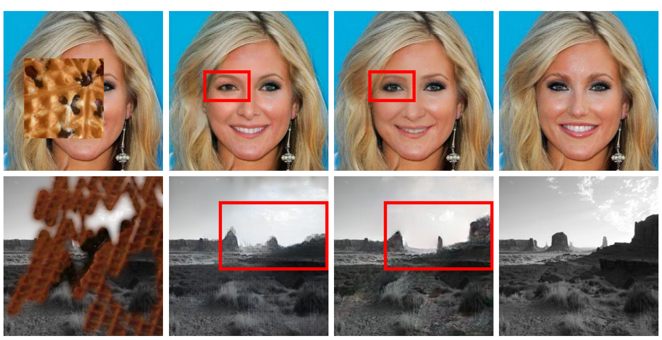
TABLE III
QUANTITATIVE EVALUATIONS ON THE CELEBAMASK-HQ [42], FFHQ [43], PARIS STREETVIEW [48] AND PLACES2 [49] WITH VARIOUS CONTAMINATION PATTERNS AS INPUT. ↓ INDICATES THE LOWER THE BETTER WHILE ↑ MEANS THE HIGHER THE BETTER.

	Dataset	w/o ACA	w/o D	AFAN	Dataset	w/o HODC	w/o D	AFAN
PSNR ↑	FFHQ	26.5408	26.9736	**27.1040**	CelebAMask-HQ	27.9465	27.4748	**28.2603**
	Paris StreetView	25.3473	26.7190	**26.9927**	Places2	26.5374	25.9581	**26.7409**
SSIM ↑	FFHQ	0.9033	0.9087	**0.9124**	CelebAMask-HQ	0.9201	0.9263	**0.9387**
	Paris StreetView	0.8613	0.8700	**0.8724**	Places2	0.8716	0.8857	**0.8983**
$\ell_1(\%)$ ↓	FFHQ	3.7346	2.2184	**2.1642**	CelebAMask-HQ	1.9305	2.0953	**1.8316**
	Paris StreetView	3.8723	2.9211	**2.8544**	Places2	2.2062	2.2637	**2.1702**
LPIPS ↓	FFHQ	0.0760	0.0504	**0.0459**	CelebAMask-HQ	0.0457	0.0486	**0.0411**
	Paris StreetView	0.0924	0.0813	**0.0805**	Places2	0.0779	0.0842	**0.0722**
FID ↓	FFHQ	12.8712	11.3538	**10.3784**	CelebAMask-HQ	9.2674	10.6353	**8.4829**
	Paris StreetView	39.3578	36.3674	**34.9745**	Places2	22.0278	22.3898	**20.5134**



(a) Input (b) w/o ACA (c) AFAN (d) GT

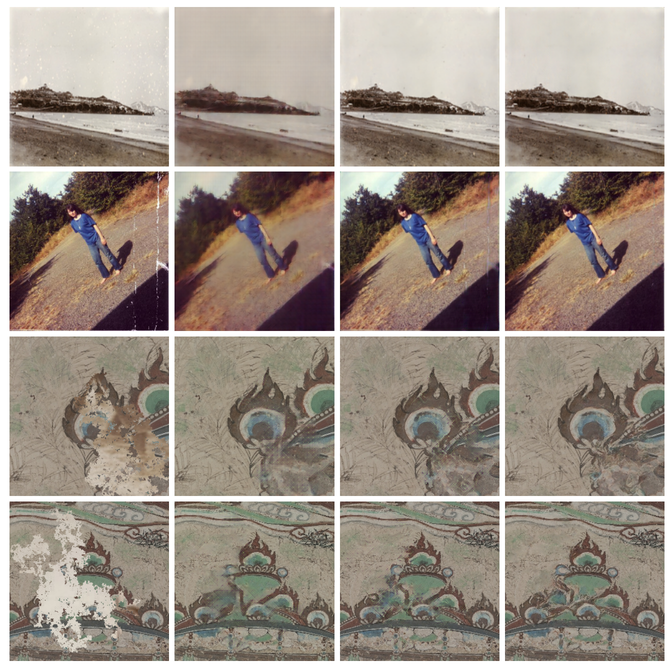
Fig. 13. Ablation study of the ACA strategy. The experiment is conducted on FFHQ [43] and Paris StreetView [48]. w/o ACA refers to the configuration where the AFAN model is trained without employing the ACA scheme.



(a) Input (b) w/o HODC (c) AFAN (d) GT

Fig. 14. Ablation study of the HODC strategy. The experiment is conducted on CelebAMask-HQ [42] and Places2 [49]. w/o HODC refers to the configuration where the AFAN model is trained without HODC layers.

variants and the full AFAN are demonstrated in Table III. The results indicate that all variant models underperformed compared to the full model. Specifically, a comparison of columns (b) and (c) in Fig. 12 shows that the proposed mask region perception strategy significantly reduces the presence of contaminant artifacts. Fig. 13 illustrates that ACA plays a crucial role in improving the precision of local feature identification while preserving rich detail. Similarly, the HODC module, leveraging edge features calculated by the Scharr operator, improves the expression of fine details. Its effectiveness is further validated by the visual results presented in Fig.14.



(a) Input (b) VCNet (c) OmniNet (d) Ours

Fig. 15. Comparison with the state-of-the-art on old photos and mural painting.

E. Application

Fig. 7 and Fig. 8 demonstrate the effectiveness of AFAN in tasks such as graffiti removal (e.g., text-like contamination patterns). Additionally, we extend AFAN to applications like old photo and mural restoration, where defects such as scratches and blemishes, which lack mask priors, require blind image inpainting techniques for accurate removal and completion. Fig. 15 shows a qualitative comparison between AFAN and state-of-the-art blind image inpainting models. The results from VCNet and OmniNet exhibit blurring artifacts and fail to completely remove scratches. In contrast, our model generates more realistic structures and preserves richer details, highlighting its superior performance in such restoration tasks.

V. Conclusion

This paper presents AFAN, a robust blind inpainting framework that exhibits significant restoration capabilities across diverse benchmark datasets. The framework leverages an adversarial training strategy, incorporating forgery detection as a mask region perception mechanism. To address both global and local content features effectively, AFAN integrates adaptive contextual attention blocks, enhancing its ability to handle contextual relationships. Additionally, high-frequency omni-dimensional dynamic convolution is implemented to capture more texture details, contributing to more realistic and detailed reconstructions. Comprehensive evaluations on various benchmark datasets demonstrate that AFAN achieves superior results in blind image inpainting for various contamination. The proposed AFAN excels in content reconstruction without relying on mask priors, expanding its applicability to more realistic scenarios. Additionally, the ACA and HODC modules 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 Workshop, 2023, pp. 1251–1260. 1, 3, 6, 7, 8, 9

[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] 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, 9

[9] 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

[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, 9

[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] J. Wang, G. Pan, D. Sun, and J. Zhang, "Chinese character inpainting with contextual semantic constraints," in ACM MM, 2021, pp. 1829–1837. 2

[35] S. Zhu, H. Xue, N. Nie, C. Zhu, H. Liu, and P. Fang, "Reproducing the past: A dataset for benchmarking inscription restoration," in ACM MM, 2024, pp. 7714–7723. 2

[36] J. Wu, X. Li, C. Si, S. Zhou, J. Yang, J. Zhang, Y. Li, K. Chen, Y. Tong, Z. Liu et al., "Towards language-driven video inpainting via multimodal large language models," in CVPR, 2024, pp. 12 501–12 511. 2

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

[38] 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

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

[40] 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

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

[42] 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, 8, 9, 10

[43] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in CVPR, 2019, pp. 4401–4410. 6, 8, 9, 10

[44] 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. 6

[45] 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. 6

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

[47] H. Scharr, "Optimal operators in digital image processing," Ph.D. dissertation, 2000. 6

[48] 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, 10

[49] 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, 10

[50] 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, 8

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

[52] 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. 8

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014. 8

[54] 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. 9

[55] 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. 9

Supplementary Materials

A. Qualitative Evaluations

To validate the inpainting performance, Fig. 1 presents comparative results from a variety of methods across an extended set of examples. The predicted results align with the descriptions provided in Section IV-B, demonstrating that our method consistently produces more reliable and high-quality inpainting results.

B. HODC Detail

Section III-D explains how HODC extends upon omni-dimensional dynamic convolution (ODC) by incorporating edge features to enhance the representation of fine details. For edge detection, several filtering operators (e.g., Scharr, Laplace, Canny, Sobel, Prewitt, Roberts) are commonly employed to highlight contours and details. Fig. 2 illustrates the edge features of these operators, as well as their fusion results with the original image. Notably, Fig. 2(b-f) demonstrates the introduction of extraneous edge artifacts, while the Scharr operation in Fig. 2(g) strikes an optimal balance between detail augmentation and structural fidelity. Consequently, the HODC employs edge features calculated by the Scharr operator to enhance the expression of details.

C. Inpainting Detail

To analyze the image reconstruction process, we visualized the feature maps generated by each component of the generator G, labeled as $[C_1, C_2, ..., C_8]$. Fig. 3 presents the dynamic visualization process, which includes examples of both high-quality face images and scene images, illustrating how the generator handles diverse visual content.

Specifically, the encoder, consisting of components $[C_1, ..., C_4]$, systematically reduces the spatial dimensions of the images. This process primarily extracts and condenses contextual information surrounding the contaminated regions within the images. Additionally, with the integration of the proposed HODC and ACA modules, the encoder captures and intricately processes features specifically related to the contaminated regions, as illustrated in Fig. 3(b-e).

Following this targeted feature extraction, the decoder, comprised of components $[C_5, ..., C_8]$, utilizes the refined features processed by the encoder to reconstruct the complete image. This reconstruction is achieved through a sequence of HODC layers and ACA blocks, designed not only to rebuild the image but also to predict and fill in the contaminated parts, effectively restoring the image to its intended state. Fig. 3(f-i) displays the reconstruction process within the decoder.

References

[1] 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. 2

[2] 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. 2

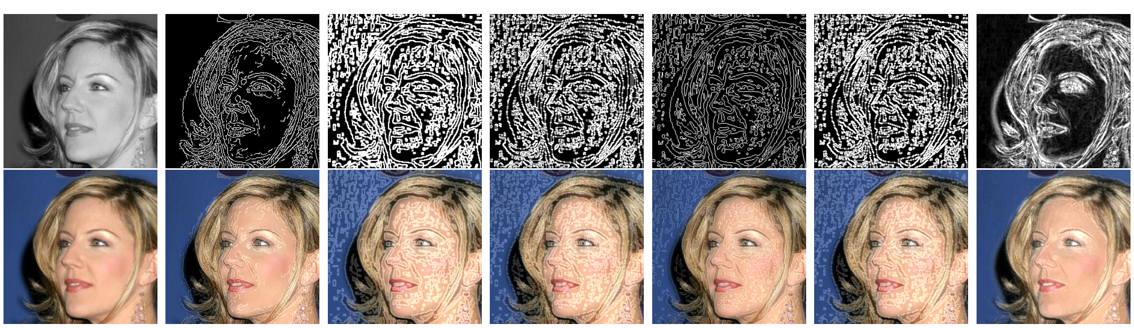
[3] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in CVPR, 2022, pp. 10 758–10 768. 2

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

[5] 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. 2

(a) Input (b) VCNet [1] (c) TransHAE [2] (d) MAT [3] (e) OmniNet [4] (f) Ours($\cancel{D}$) (g) Ours (h) GT

Fig. 1. Comparison with the state-of-the-art. These images come from datasets listed in Section 4.1 with various contamination patterns. Ours($\cancel{D}$) refers to the configuration where the AFAN model is trained without employing the mask supervision strategy denoted as D.



(a) GT (b) Canny (c) Laplacian (d) Prewitt (e) Roberts (f) Sobel (g) Scharr

Fig. 2. Comparison of different edge detection filters. The first row shows the edge features of the corresponding operator, and the second row shows the fusion result with the original image. The experiment is conducted on the CelebAMask-HQ [5] dataset.

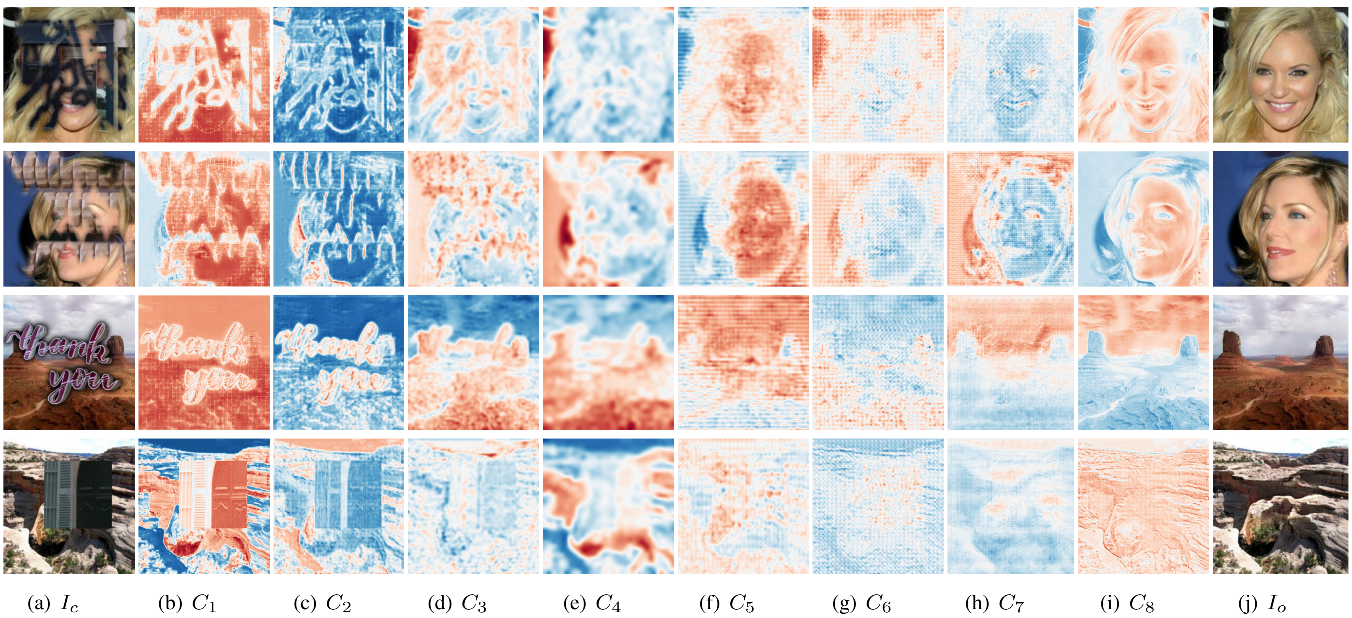


Fig. 3. Feature map visualization of the generator. C represents the 8 transformer-style components, I_c is the contaminated image and I_o is the inpainted image. $C_1 - C_4$ are encoder components and C_5–C_8 are decoder components.