Blind Image Inpainting via Omni-dimensional Gated Attention and Wavelet Queries

Shruti S. Phutke, Ashutosh Kulkarni, Santosh Kumar Vipparthi, Subrahmanyam Murala
Computer Vision and Pattern Recognition Lab
Indian Institute of Technology Ropar, Rupnagar, Punjab
{2018eez0019, ashutosh.20eez0008, skvipparthi, subbumurala}@iitrpr.ac.in

Abstract

Blind image inpainting is a crucial restoration task that does not demand additional mask information to restore the corrupted regions. Yet, it is a very less explored research area due to the difficulty in discriminating between corrupted and valid regions. There exist very few approaches for blind image inpainting which sometimes fail at producing plausible inpainted images. Since they follow a common practice of predicting the corrupted regions and then inpaint them. To skip the corrupted region prediction step and obtain better results, in this work, we propose a novel end-to-end architecture for blind image inpainting consisting of wavelet query multi-head attention transformer block and the omni-dimensional gated attention. The proposed wavelet query multi-head attention in the transformer block provides encoder features via processed wavelet coefficients as query to the multi-head attention. Further, the proposed omni-dimensional gated attention effectively provides all dimensional attentive features from the encoder to the respective decoder. Our proposed approach is compared numerically and visually with existing state-of-the-art methods for blind image inpainting on different standard datasets. The comparative and ablation studies prove the effectiveness of the proposed approach for blind image inpainting. The testing code is available at : https://github.com/shrutiphutke/Blind_Omni_Wav_Net

1. Introduction

Image inpainting is a widely used technique in the field of image processing and restoration. This method involves filling in missing or corrupted regions of an image to restore it to its original form. Typically, image inpainting methods require information about the corrupted regions in the form of masks to guide the restoration process. These methods are known as non-blind image inpainting methods. However, in many real-world applications, such as photo editing, unwanted object removal, mesh-face verification, etc., it is often difficult to obtain masks for guidance. This has led to the development of a new technique called blind image inpainting, which does not require any prior knowledge about the corrupted regions or masks to perform image restoration.

Figure 1. Sample visual results of the proposed method for blind image inpainting on different datasets (from left to right CelebA-HQ [9], FFHQ [10], ParisSV [5], and Places2 [41]).

As the corrupted regions in the images are arbitrary (in size and shape), inpainting becomes an ill-posed problem. The deep learning-based architectures nowadays effectively inpaint the corrupted images [6, 8, 11, 15, 16, 19, 23, 24, 30, 35]. These methods use different approaches such as utilizing multiple cues to inpaint the image [35], generative adversarial networks (GANs) inversion [30], pseudo-decoder [23], super-resolution [11], transformers [6, 15], *etc.* Though, these methods are excellent in producing the faithful inpainting outcomes, they generally provide the mask as input to guide the network.

With regard to blind image inpainting, considering that there is no knowledge of where the corruptions lie (masks), image inpainting becomes a difficult task. Researchers divided the task of blind image inpainting into two sub-tasks *i.e.* (a) mask prediction and (b) image inpainting [29, 31].

In [31], authors proposed a mask prediction network to predict the visually disturbed regions followed by the robust image inpainting network. Also, [29] used the transformer-based network for mask prediction followed by top-down refinement inpainting network. In both of these approaches, the final inpainting task is heavily dependent upon the earlier mask prediction network prior to beginning the inpainting process. There may be a degradation in the performance of inpainting when proper mask prediction is not achieved. For overcoming the above limitations of mask prediction-based blind image inpainting, a hybrid transformer encoder with cross-layer dissimilarity prompt (CDP) and convolution neural network (CNN) decoder is proposed in [40] for the identification of contamination in an input image. There may be a possibility that the generation of an attentive soft mask for further inpainting could introduce inconsistencies in the inpainted results where a CNN decoder is being used. Since, it may fail to take into account the maximum dependencies from the input mask attentive feature maps.

Despite the apparent differences between stage-wise mask prediction followed by inpainting [29, 31] and intermediate attentive mask guidance inpainting [40], both approaches are nearly identical. Since, in that the inpainting task is directly or indirectly based on the identification of inconsistent regions. Also, when an image with large corrupted regions is provided as input to these methods [29, 31, 40], they fail at inpainting the globally consistent image. In this work, we propose an end-to-end training approach independent of any identification of masked and non-masked regions for blind image inpainting.

The transformers are well known for their ability to exploit long-range dependencies. With this ability, transformers have shown better convergence in numerous applications of image restoration [6, 15, 19, 28, 29, 39] including image inpainting. Further, the queries are the inputs to the transformer multi-head attention for which the attention is calculated. Providing appropriate queries to the transformer block may further enhance their convergence capability. With this assumption, in this work, we propose a wavelet query-based multi-head attention mechanism in the transformer block. The processed wavelet coefficients will provide less degraded information as a query to the multi-head attention mechanism. Also, to forward the encoder features to the respective decoder, we propose a gated omni-dimensional attention block. This block provides the all dimensional attentive information to the features which may help the network for efficient reconstruction. The contributions of our work are:

- An end-to-end transformer based architecture is proposed for blind image inpainting.
- A novel wavelet query multi-head attention mechanism is introduced in the transformer block.
- A omni-dimensional gated attention mechanism is proposed to forward different dimensional attentive features from encoder to respective decoder for effective reconstruction of inpainted image.

Our proposed approach achieves remarkable performance improvement as compared to existing state-of-the-art blind image inpainting methods. The sample inputs and outputs of the proposed approach are provided in Figure 1.

2. Related Work

2.1. Image Inpainting

Previously, image inpainting methods [1, 4, 12] primarily relied on conventional patch, exemplar, and diffusion-based techniques, but these often resulted in semantically inconsistent results. Later on, novel works on image inpainting [6, 11, 15, 23, 30, 35] were proposed, which utilized learning-based approaches and made significant advancements in this field. Methods relying on two stage architectures [6, 16, 22, 24, 36, 37] resulted in remarkable image inpainting outcomes. These two stage architectures mainly comprise of an initial stage, which produces coarse output, and a second stage, which generates the finer results. Although the two-stage architectures produce believable outcomes, their inter-stage dependency hinders them from generating more detailed and refined inpainted results. To mitigate this challenge, improvements in the form of single stage inpainting architectures with less computational complexity were introduced [6, 14, 15, 18, 19, 23, 38] that perform better than the two stage architectures. By leveraging the mask information, researchers proposed mask aware convolution layers [18, 37], contextual attention [14], and transformers [15, 38] for image inpainting. Further, additional information such as structural or edge priors have been utilized in [21, 26] for image inpainting. These above methods utilize the mask information for inpainting which limit its performance where mask information is not available. Therefore, researchers introduced the inpainting architectures without any mask information named as blind image inpainting.

2.2. Blind Image Inpainting

Cai *et al*. [2] proposed the first blind image inpainting with end-to-end CNN based architecture for direct learning to identify the corrupted regions and recover them. Later, Liu *et al*. [20] proposed a residual learning based approach with the horizontal and vertical gradients to generate the detailed clear image. Prior to these works, Xie *et al*. [34] utilized the sparse auto-encoder for image denoising and blind image inpainting. Similarly, in [25] Ren *et al*. proposed Shepard convolutional network for image denoising and blind image inpainting. These approaches consider simple contaminations like text imposed on images or images with

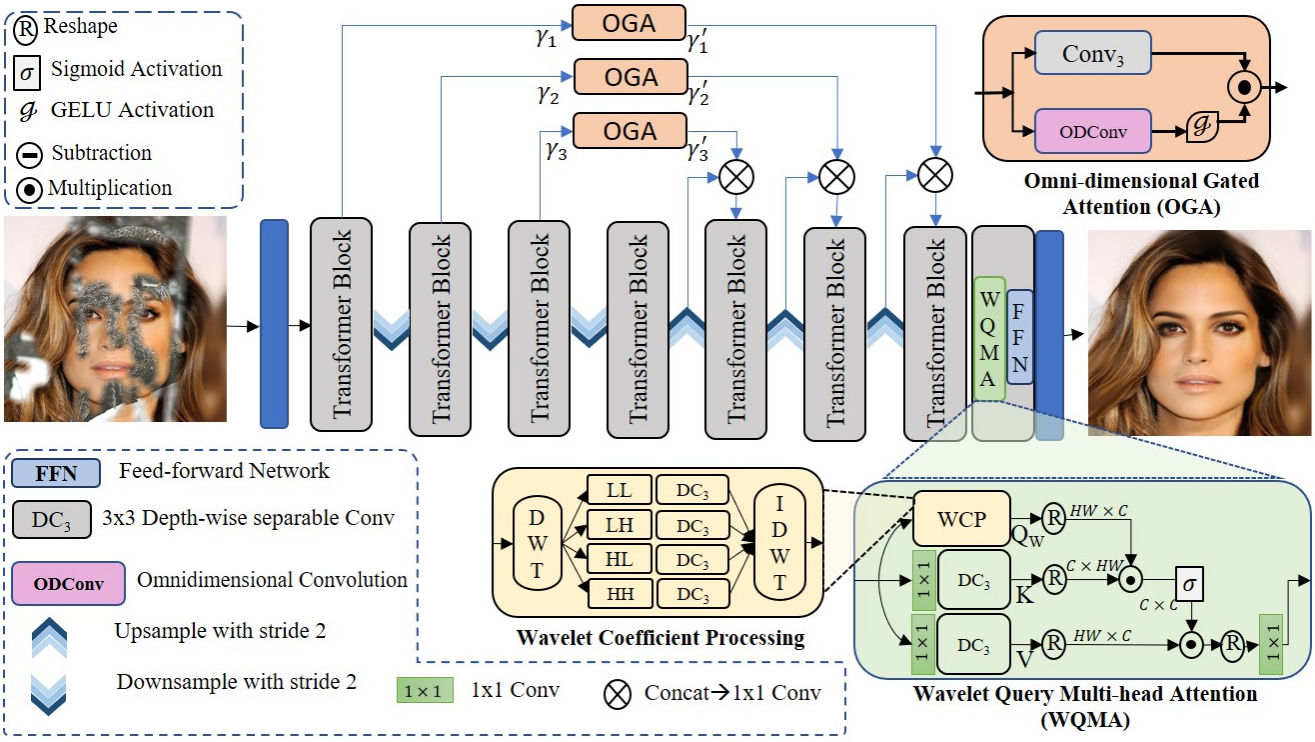

Figure 2. Overview of the proposed architecture for blind image inpainting. The architecture includes of transformer block consisting of the proposed wavelet query-based multi-head attention for providing prominent information as a query. Further, omni-dimensional gated attention is proposed in order to forward efficient attentive features from encoder to the respective decoder.

some part appended from other images of masks with thin size. Due to this, these methods may fail at inpainting the images with larger contaminations.

For considering complex contaminations, Wang *et al.* [31] proposed two-stage architecture for blind image inpainting. [31] put forward a consistency network that initially recognizes the regions to inpaint through the predicted masks, and then leverages the inpainting network. Further, Wang *et al.* [29] predicted the masks by including the contextual coherence and additional frequency modality as inputs, after which, landmark prediction and finally inpainting is performed on facial images. However, both [31] and [29] rely on mask prediction followed by inpainting by using two-stage architectures. Such approaches are susceptible to mask prediction error, further leading to undesired image inpainting results. With this knowledge, Zhao *et al.* [40] observed that, in blind image inpainting, there exists a heavy correlation between the differentiation of contaminated and valid regions and mask prediction. With this assumption, Zhao *et al.* [40] proposed a single-stage hybrid encoder-decoder network for blind image inpainting. In order to capture global context, the transformer encoder is used [40] and the CNN decoder is used to revamp the contaminations.

2.3. Transformers for Image Restoration

Transformers have gained popularity over CNNs for various image restoration tasks such as image denoising [17], deraining [33], deblurring [32, 39], *etc.* due to their ability to capture long-range dependencies. They have been used in computer vision through Vision Transformers (ViT) [7], which employ flattened patches of images while training. Image processing transformers have also been used for low-level vision tasks, as demonstrated by [3], which demonstrated how pre-training on large datasets can improve performance. UFormer [32] uses a U-Net like structure with windowed multi-head self-attention based transformers for image restoration tasks. Further, Zamir *et al.* [39] introduced efficient transformers for various image restoration tasks.

3. Proposed Method

For the task of blind image inpainting, corrupted input $C \in \mathbb{R}^{m,n,3}$ with m, n spatial dimensions is generated as:

$$C = G_t \odot (1 - \mathbb{G}(Mask)) + Noise \odot \mathbb{G}(Mask) \quad (1)$$

where, $G_t \in \mathbb{R}^{m,n,3}$ is the ground-truth image, $Noise \in \mathbb{R}^{m,n,3}$ is the noisy input, $Mask$ is the input which decides

the locations to be corrupted with 1 as corrupted regions and 0 as valid regions, and $\mathbb{G}(\cdot)$ is Gaussian smoothing which is applied on $Mask$. As, this work is mainly focused on blind image inpainting, the proposed network does not utilize the $Mask$ information for inpainting the image.

In this work, we propose a single-stage end-to-end transformer architecture for blind image inpainting (*see Figure 2*). Here, we propose two major components namely: (a) wavelet query multi-head attention mechanism in transformer: *to provide processed query as input to the multi-head attention*, and (b) omnidimensional gated attention: for *providing all dimensional attentive features* in order to achieve a plausible outcome. In this section, we will first give a detailed exposition of the proposed architecture for blind image inpainting and then we detail the proposed modules.

Overview of proposed transformer based architecture with the wavelet query multi-head attention (WQMA) and omnidimensional gated attention (OGA) is shown in Figure 2. To convert input image into feature space, we first apply the convolution layer. These convolved features are processed through three successive transformer blocks followed by down-sampler. The input with spatial size m, n is then converted into $\frac{m}{8}, \frac{n}{8}$ sized feature maps at 4^{th} transformer block. In this transformer block, we propose *a wavelet query multi-head attention (WQMA) to provide processed features as a query to the multi-head attention.* These feature maps are then forwarded again to the successive transformer blocks but now these blocks are followed with an up-sampler to come up with the actual spatial dimension (m, n) at the last stage. Here, in the decoding stage, we apply the proposed omni-dimensional gated attention (OGA) on encoder features while giving a skip connection from the respective encoder to the subsequent decoder level. *The OGA helps the network to provide multidimensional attentive features to the decoder for effective reconstruction.* The structure of the transformer block consists of the proposed WQMA and a feed-forward network [39]. Finally, we again apply a convolution layer to generate final output O.

3.1. Wavelet Query Multi-head Attention

In the existing transformer approach [39], generally the query, key, and values are considered from the same input without any separate processing to generate them. In a transformer block, the query is used to which attention is calculated and the key is from which the attention is calculated. So, here query plays an important role in overall multi-head attention for which attention is calculated. Providing effective features as a query may help the transformer block to further improve its performance. The contaminations in the inputs for a blind image inpainting task are considered as the noise appended on top of the clear image. Wavelets are well known for the task of image denoising where each of the decomposed wavelet coefficients is processed separately to reduce the noise. The wavelet-based attention mechanism is proposed in [42] for the task of image classification where the attention mechanism is applied in wavelet coefficient space. In the case of image inpainting, the input image has some corrupted regions present in it. Directly applying the attention in wavelet coefficient space may consider the corrupted regions also. Since the wavelet coefficient space also has corrupted regions in it. In order to avoid forwarding the noisy wavelet coefficients, we propose the processing of each wavelet coefficient. Further, the multi-head attention mechanism plays an essential role in capturing the long-term dependencies in the transformer block. The 2D wavelet coefficients are first calculated using forward discrete wavelet transform (DWT) as:

$$LL, LH, HL, HH = DWT(\mathbb{F}_{in}) \tag{2}$$

where, LL, LH, HL, and HH are approximate, horizontal, vertical, and diagonal coefficients respectively of input feature maps $\mathbb{F}_{in}$ calculated using DWT. Each of the coefficients is separately processed as:

$$\begin{aligned} LL' = \psi_a(LL); LH' = \psi_h(LH) \\ HL' = \psi_v(HL); HH' = \psi_d(HH) \end{aligned} \tag{3}$$

where, ψ is depth-wise separable convolution with kernel size 3×3. Further, these processed wavelet coefficients are utilized to form the output feature map by passing them through the inverse discrete wavelet transform (see Wavelet Coefficient Processing block in Figure 2). These processed wavelet coefficients are considered as the queries (Q_W) to the multi-head attention. This may help the network to calculate the attention with less effect of contaminations. The overall attention using wavelet queries is calculated as:

$$Attention(\mathbb{F}_{in}) = \sigma\left(\frac{Q_W K^T}{\sqrt{d}}\right) V \tag{4}$$

where, $K = C_1(\psi(\mathbb{F}_{in}))$, $V = C_1(\psi(\mathbb{F}_{in}))$, C_1 is convolution with kernel size 1×1. This proposed approach helps the network to effectively capture long-term dependencies with the minimum effect of corrupted regions.

3.2. Omni-dimensional Gated Attention

In order to forward the encoder features to the respective decoder, we propose an omni-dimensional gated attention mechanism. This attention mechanism is given as:

$$\gamma_i' = C_3(\gamma_i) \odot \mathcal{G}(ODC_3(\gamma_i)) \tag{5}$$

where, γ_i are the encoder features with $i \in (1, 2, 3)$, C_3 is convolution with kernel size 3×3, $\mathcal{G}$ is a GELU activation

function, ODC is omni-dimensional convolution with kernel size 3×3. This omni-dimensional gated attention provides the weighted feature from four different dimensions to the input encoder features.

The omni-dimensional convolution is a dynamic convolution that considers all the different dimensions of the input feature maps. Here, the omni-dimensional refers to the four different dimensions *i.e*. spatial, channel, filter, and kernel-wise attention. Let, for a dynamic convolution there are n different convolutional kernels, each of the kernels has the spatial dimension $k \times k$, the number of input channels is c_{in}, and the number of output filters is c_{out}. Input (γ_i) to the ODC is first processed through a global average pooling operation followed by a fully connected layer and the ReLU activation function. These processed $1D$ features are used to generate different attentions like (i) spatial attention (α_s) of size $k \times k$ to the spatial dimension of convolution kernel, (ii) channel attention (α_c) of size $1 \times 1 \times c_{in}$ to the input channels c_{in}, (iii) filter attention (α_f) of size $1 \times 1 \times c_{out}$ to the output number of filters c_{out}, and (iv) kernel attention (α_w) to the n dynamic convolution kernels. These attentions are calculated by applying a fully connected layer (to generate the required dimension) followed by the Sigmoid activation function. The output of ODC is formulated as:

$$Y = \left(\sum_{i=1}^{n} \alpha_{w_i} \odot \alpha_{f_i} \odot \alpha_{c_i} \odot \alpha_{s_i} \odot W_i \right) * \gamma_i \quad (6)$$

where, α_{w_i} is the attention applied to i_{th} convolution kernel, α_f is the attention applied to the c_{out} convolution filters, α_c is the attention applied to the c_{in} convolution filters, and α_s is attention applied to spatial dimension $k \times k$ of convolution filter [13]. This ODA provides the network with the ability to learn attentive features from all the dimensions, unlike existing only spatial or channel-wise attentions.

4. Experiments and Results Discussion

In this section, we will discuss different experimental datasets, evaluation metrics, and quantitative and qualitative results of the proposed and existing state-of-the-art methods.

4.1. Datasets and Evaluation Metrics

For blind image inpainting, we use four datasets: FFHQ [10], CelebA-HQ [9], Places2 [41], and Paris Street View(ParisSV) [5]. The comparative analysis for blind image inpainting is done with VCNet [31], TransCNNHAE [40] (blind inpainting methods) and CTSDG [8] (non-blind inpainting method as provided in [40]). For fair comparison we have compared methods with publicly available source codes on all the blind/non-blind image inpainting datasets.

Table 1. Ablation study on different configurations of the proposed network on ParisSV dataset for blind image inpainting. (Note: ↑- Higher is better, ↓- Lower is better).

Network Configuration	PSNR↑	SSIM↑	$L_1 \downarrow$	FID ↓
TransCNNHAE [40]	26.72	0.896	0.0352	41.50
Q_W, K_W, V_W	26.89	0.885	0.0347	46.63
$Q_W, K_W, V_W + OGA$	27.50	0.901	0.0324	43.11
Q_W, K, V	27.05	0.898	0.0328	44.32
$Q_W, K, V + OGA$	**27.81**	**0.905**	**0.0301**	**40.646**

For quantitative results comparison of the proposed method and existing state-of-the-art methods on blind image inpainting, peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), mean L_1 error and Fréchet inception distance (FID) metrics are used.

4.2. Implementation and Training Details

To train the proposed blind image inpainting approach, we use AdamW optimizer with $3e^{-4}$ learning rate which is gradually reduced with the cosine annealing strategy. We train the proposed network using the L_1 loss. Also, to guide the network for textural and structural information by extracting effective features, the perceptual loss (L_P) is calculated between the deep feature maps of the ground-truth and inpainted images by passing them through the pre-trained VGG16 model [27] as:

$$L_P = \sum_{s=1}^{S} \left(\|\phi_s(G_t) - \phi_s(O)\|_1 \right) \quad (7)$$

where, G_t is ground-truth, O is the output, ϕ_s are the feature maps ($s \in (1, S)$) of the VGG16 model. The edge loss (L_e) is also considered to focus on edge enhancement while training. The edge loss with sobel operator $\mathbb{S}$ is formulated as:

$$L_e = \|\mathbb{S}(G_t) - \mathbb{S}(O)\|_1 \quad (8)$$

For structurally consistent output generation we utilized the structural similarity loss (L_S), given as:

$$L_S = 1 - SSIM(O) \quad (9)$$

where, $SSIM$ is structural similarity index metric. So the overall loss to train the network is given as:

$$L_T = \lambda_1 L_1 + \lambda_P L_P + \lambda_e L_e + \lambda_S L_S \quad (10)$$

where, λ_{Loss} is the weight assigned to respective $Loss$ function which is verified experimentally as: $\lambda_1 = 10$, $\lambda_P = 0.6$, $\lambda_e = 0.4$, $\lambda_S = 0.5$.

4.3. Ablation Study

To determine the design choices of the network for blind image inpainting, we performed various experiments on

Figure 3. Qualitative result analysis of ablation study on different configurations of the proposed network for blind image inpainting.

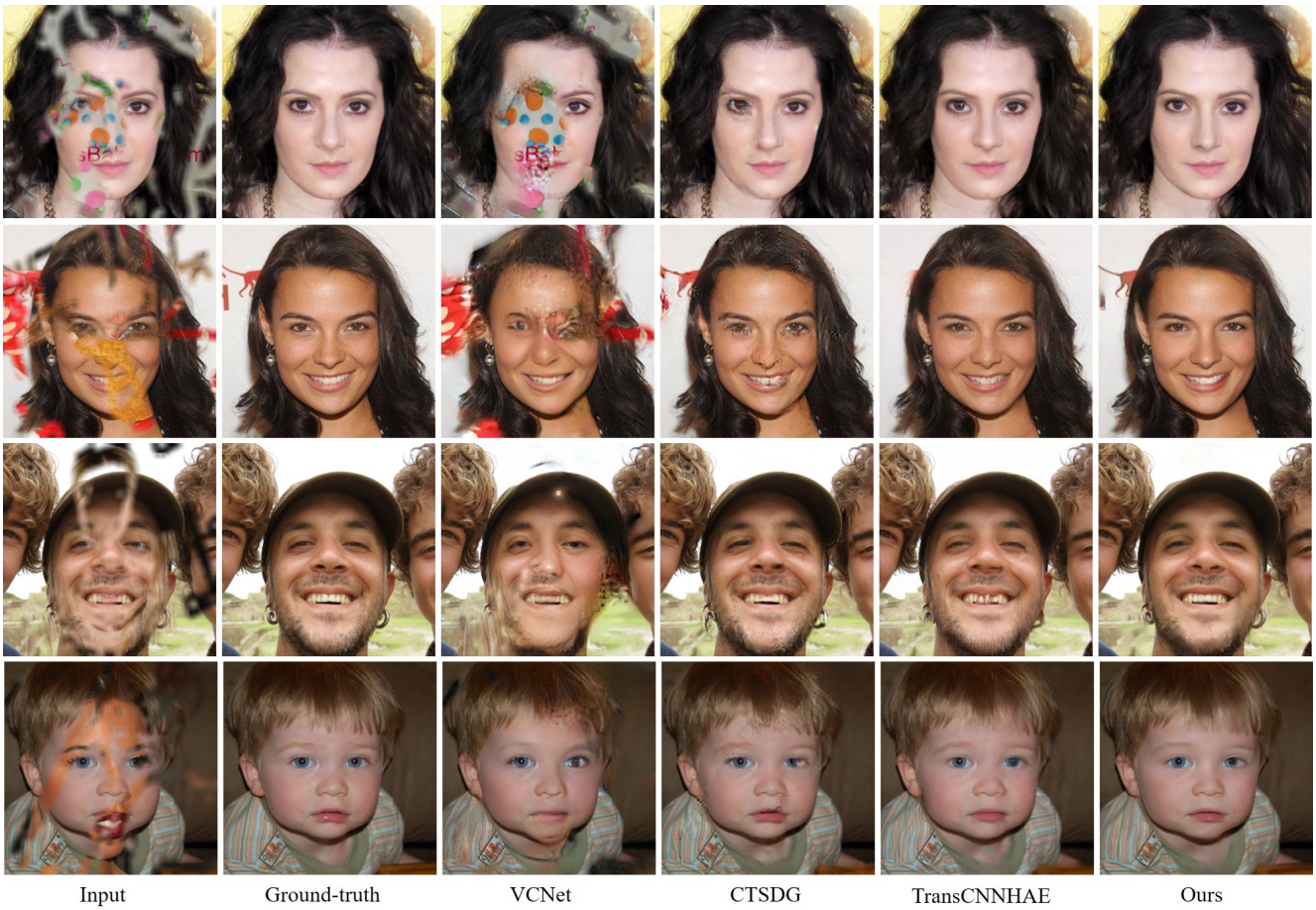

Figure 4. Qualitative results comparison of the proposed method (Ours) with existing state-of-the-art methods (VCNet [31], CTSDG [8], TransCNNHAE [40]) on Celeb (first two rows) and FFHQ (last two rows) dataset for blind image inpainting.

the Paris_SV dataset. How the each of proposed modules led to performance improvement is discussed in this section.

4.3.1 Effect of the wavelet-based query to multi-head attention

Wavelet base attention mechanism in transformer block has proved its efficiency for the image classification task [42]. With this motivation, at first, we aimed to provide wavelet query (Q_W), keys (K_W), and values (K_W) to the multi-head attention. For comparison purpose, we considered the existing best blind image inpainting method (TransCNNHAE [40]). The results improved in terms of PSNR, SSIM, and L_1 error. But there was no improvement in FID due to structural inconsistencies. Further, we evaluated the importance of providing wavelet processed query

Figure 5. Qualitative results comparison of the proposed method (Ours) with existing state-of-the-art methods (VCNet [31], CTSDG [8], TransCNNHAE [40]) on Paris_SV (first two rows) and Places2 (last two rows) datasets for blind image inpainting.

Table 2. Comparison of the proposed method (Ours) and existing state-of-the-art methods for blind image inpainting (↑- Higher is better, ↓- Lower is better).

Metric	Dataset	VCNet [31]	CTSDG [8]	TransCNNHAE [40]	Ours
PSNR ↑	CelebA-HQ	25.59	26.94	27.71	**28.21**
	FFHQ	23.62	24.62	27.05	**28.19**
	ParisSV	23.62	26.08	26.72	**27.81**
	Places2	24.09	26.05	26.87	**27.55**
SSIM ↑	CelebA-HQ	0.874	0.934	0.949	**0.951**
	FFHQ	0.861	0.935	0.941	**0.952**
	ParisSV	0.824	0.861	0.896	**0.905**
	Places2	0.869	0.905	0.910	**0.918**
L_1 ↓	CelebA-HQ	0.0396	0.0318	0.0250	**0.0221**
	FFHQ	0.0482	0.0392	0.0281	**0.0234**
	ParisSV	0.0527	0.0412	0.0352	**0.0301**
	Places2	0.0429	0.0308	0.0261	**0.0231**
FID ↓	CelebA-HQ	9.275	8.561	7.251	**7.235**
	FFHQ	10.148	9.586	9.424	**8.639**
	ParisSV	64.215	43.015	41.505	**40.646**
	Places2	28.821	18.685	17.640	**17.521**

Figure 6. Qualitative results comparison of the proposed method (Ours) with existing state-of-the-art method (TransCNNHAE [40]) on unseen patterns.

Table 3. Computational complexity analysis (the **best** and second best are shown in **bold** and underline).

Method	Parameters (M) ↓	FLOPs (G) ↓
VCNet [31]	3.79	65.25
CTSDG [8]	52.14	53.38
TransCNNHAE [40]	**2.75**	19.71
Ours	3.24	**16.61**

only to the multi-head attention with a combination of Q_W, K, V which resulted in better convergence as compared to Q_W, K_W, V_W (see row 2 and 4 of Table 1)

4.3.2 Effect of omni-dimensional gated attention

Further, to help the network for better reconstruction and structural information, we proposed omni-dimensional gated attention (OGA). The experiments are carried out with both the above discussed wavelet conditions *i.e.* $Q_W, K_W, V_W + OGA$ and $Q_W, K, V + OGA$ to verify the effectiveness of both the proposed modules. The inclusion of the proposed OGA to forward the encoder features to the respective decoder performed well by imroving in terms of PSNR and SSIM. Along with these parameters improvement, there is a lot of improvement in the FID value (see Table 1).

Overall, our proposed modules ($Q_W, K, V + OGA$) effectively help the network with improved performance for the task of blind image inpainting. Also, the visual results of the ablation study are provided in Figure 3.

4.4. Blind Image Inpainting Results Analysis

For the task of blind image inapinting, we considered four different datasets covering large variety of cases like natural places scenes, facial images. The comparison in terms of PSNR, SSIM L_1 error and FID is provided in Table 2. Along with state-of-the-art blind image inpainting methods [31, 40] ([40] is retrained on respective datasets as per the configurations provided due to unavailability of pre-trained checkpoints), we considered the existing non-blind image inpainting method [8] with best performance (*as provided in [40]*). Since, it is worth to note that, the existing non-blind method may not work feasibly for blind image inpainting task, we provided the ground-truth masks as inputs to these methods as suggested in [40]. From Table 2, it is clear that the proposed approach for blind image inpainting performs remarkably as compared to state-of-the-art blind and non-blind methods.

The visual results comparison for blind image inpainting is provided in Figure 4 and 5. When compared qualitatively, our proposed method generates comparatively plausible results on all the datasets for blind image inpainting.

The computational complexity comparison of the proposed approach and existing methods is given in Table 3 in terms of the number of trainable parameters and the number of floating point operations (FLOPs). Although moderately complex in terms of the number of trainable parameters, our proposed approach has less complexity in terms of the number of FLOPs as compared to state-of-the-arts.

4.5. Unseen Contamination Result Analysis

Here, we have evaluated the performance of our proposed approach for unseen contamination such as random scratches and text. The comparison is done with the existing state-of-the-art (TransCNNHAE [40]) for blind image inpainting. Figure 6 shows the performance of our proposed approach on unseen patterns as compared to existing approach for blind image inpainting.

5. Conclusion

This work proposes an end-to-end transformer approach for blind image inpainting. We propose the wavelet coefficient processing and providing them as a query to multi-head attention in the transformer block. Further, the gated omni-dimensional attention is proposed to forward the encoder features to the respective decoder as a skip connection. A series of ablation studies are carried out to demonstrate the feasibility of the proposed modules. The quantitative and qualitative comparison of the proposed network with existing state-of-the-art methods for blind and non-blind methods verifies the reliability of the proposed architecture for the task of blind image inpainting. Also, the performance of the proposed approach is verified for unseen contamination. This approach can be further extended for other restoration tasks such as image rain and snow removal.

References

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

[2] Nian Cai, Zhenghang Su, Zhineng Lin, Han Wang, Zhijing Yang, and Bingo Wing-Kuen Ling. Blind inpainting using the fully convolutional neural network. *The Visual Computer*, 33(2):249–261, 2017. 2

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

[4] Ding Ding, Sundaresh Ram, and Jeffrey J Rodríguez. Image inpainting using nonlocal texture matching and nonlinear filtering. *IEEE Transactions on Image Processing*, 28(4):1705–1719, 2018. 2

[5] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei Efros. What makes paris look like paris? *ACM Transactions on Graphics*, 31(4), 2012. 1, 5

[6] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022. 1, 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[8] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14134–14143, 2021. 1, 5, 6, 7, 8

[9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 1, 5

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 5

[11] Soo Ye Kim, Kfir Aberman, Nori Kanazawa, Rahul Garg, Neal Wadhwa, Huiwen Chang, Nikhil Karnad, Munchurl Kim, and Orly Liba. Zoom-to-inpaint: Image inpainting with high-frequency details. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 477–487, 2022. 1, 2

[12] Olivier Le Meur, Josselin Gautier, and Christine Guillemot. Examplar-based inpainting based on local geometry. In *2011 18th IEEE international conference on image processing*, pages 3401–3404. IEEE, 2011. 2

[13] Chao Li, Aojun Zhou, and Anbang Yao. Omni-dimensional dynamic convolution. *arXiv preprint arXiv:2209.07947*, 2022. 5

[14] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768, 2020. 2

[15] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10758–10768, 2022. 1, 2

[16] Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, and Song Wang. Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1869–1878, 2022. 1, 2

[17] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 3

[18] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 2

[19] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. Reduce information loss in transformers for pluralistic image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11347–11357, 2022. 1, 2

[20] Yang Liu, Jinshan Pan, and Zhixun Su. Deep blind image inpainting. In *International Conference on Intelligent Science and Big Data Engineering*, pages 128–141. Springer, 2019. 2

[21] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[22] Shruti S Phutke and Subrahmanyam Murala. Fasnet: Feature aggregation and sharing network for image inpainting. *IEEE Signal Processing Letters*, 29:1664–1668, 2022. 2

[23] Shruti S Phutke and Subrahmanyam Murala. Pseudo decoder guided light-weight architecture for image inpainting. *IEEE Transactions on Image Processing*, 2022. 1, 2

[24] Weize Quan, Ruisong Zhang, Yong Zhang, Zhifeng Li, Jue Wang, and Dong-Ming Yan. Image inpainting with local and global refinement. *IEEE Transactions on Image Processing*, 31:2405–2420, 2022. 1, 2

[25] Jimmy S Ren, Li Xu, Qiong Yan, and Wenxiu Sun. Shepard convolutional neural networks. *Advances in neural information processing systems*, 28, 2015. 2

[26] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019. 2

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[28] Jeya Maria Jose Valanarasu, Rajeev Yasarla, and Vishal M Patel. Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2353–2363, 2022. 2

[29] Junke Wang, Shaoxiang Chen, Zuxuan Wu, and Yu-Gang Jiang. Ft-tdr: Frequency-guided transformer and top-down refinement network for blind face inpainting. *IEEE Transactions on Multimedia*, 2022. 1, 2, 3

[30] Wentao Wang, Li Niu, Jianfu Zhang, Xue Yang, and Liqing Zhang. Dual-path image inpainting with auxiliary gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11421–11430, 2022. 1, 2

[31] Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. Vcnet: A robust approach to blind image inpainting. In *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 1, 2, 3, 5, 6, 7, 8

[32] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, June 2022. 3

[33] Jie Xiao, Xueyang Fu, Aiping Liu, Feng Wu, and Zheng-Jun Zha. Image de-raining transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[34] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25, 2012. 2

[35] Yohei Yamashita, Kodai Shimosato, and Norimichi Ukita. Boundary-aware image inpainting with multiple auxiliary cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 619–629, 2022. 1, 2

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

[37] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4471–4480, 2019. 2

[38] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021. 2

[39] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2, 3, 4

[40] Haoru Zhao, Zhaorui Gu, Bing Zheng, and Haiyong Zheng. Transcnn-hae: Transformer-cnn hybrid autoencoder for blind image inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6813–6821, 2022. 2, 3, 5, 6, 7, 8

[41] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1, 5

[42] Yufan Zhuang, Zihan Wang, Fangbo Tao, and Jingbo Shang. Waveformer: Linear-time attention with forward and backward wavelet transform. *arXiv preprint arXiv:2210.01989*, 2022. 4, 6