Nested Deformable Multi-head Attention for Facial Image Inpainting

Shruti S Phutke and Subrahmanyam Murala
CVPR Lab, Indian Institute of Technology Ropar, Punjab, INDIA
{2018eez0019, subbumurala}@iitrpr.ac.in

Abstract

Extracting adequate contextual information is an important aspect of any image inpainting method. To achieve this, ample image inpainting methods are available that aim to focus on large receptive fields. Recent advancements in the deep learning field with the introduction of transformers for image inpainting paved the way toward plausible results. Stacking multiple transformer blocks in a single layer causes the architecture to become computationally complex. In this context, we propose a novel lightweight architecture with a nested deformable attention-based transformer layer for feature fusion. The nested attention helps the network to focus on long-term dependencies from encoder and decoder features. Also, multi-head attention consisting of a deformable convolution is proposed to delve into the diverse receptive fields. With the advantage of nested and deformable attention, we propose a lightweight architecture for facial image inpainting. The results comparison on Celeb_HQ [25] dataset using known (NVIDIA) and unknown (QD-IMD) masks and Places2 [57] dataset with NVIDIA masks along with extensive ablation study prove the superiority of the proposed approach for image inpainting tasks. The code is available at: `https://github.com/shrutiphutke/NDMA_Facial_Inpainting`.

1. Introduction

Image inpainting is a perennial task of filling the holes with the most probable contents which generate a plausible outcome. The wide variety of applications such as 3D image generation, photo restoration, object removal, portrait editing, *etc.* has made image inpainting a popular computer vision task. The conventional inpainting methods [5, 12, 14] made use of textural or patch-based statistical information to inpaint images. These methods lack in generating high-level semantics and structurally plausible results.

With the advancement in convolutional neural networks (CNNs) and generative adversarial networks (GANs), various methods are proposed for image inpainting which generate faithful results [47, 46, 49, 41, 30, 31, 32]. The main aspect of the inpainting task is to extract the relevant contextual information from various receptive fields. The pioneer works with deep learning approach achieve the contextual information extraction by utilizing various phenomena such as recurrent feature processing [19], hyper-graphs [38], contextual reconstruction [52], *etc.* In [36], the authors proposed a Fast Fourier Convolution-based feature encoding for global receptive fields. For the image inpainting task, one can have ample data for training by corrupting the clean images with different masks. With this advent, numerous methods are proposed and deliver promising inpainting outcomes. Still, they lack in producing realistic outcomes due to distorted structures and blurriness. Also, some methods are proposed with prior knowledge to produce a faithful outcome.

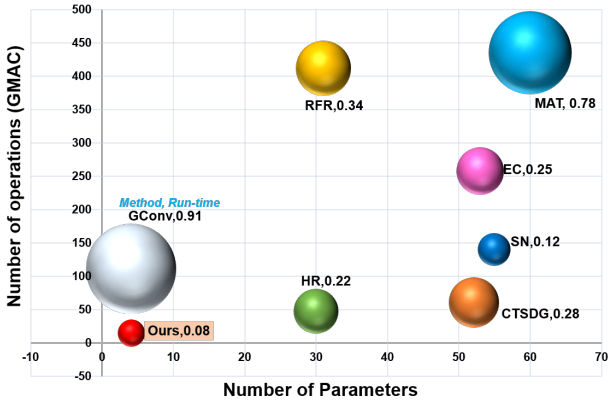

Figure 1. Comparison of the proposed method (ours) with existing methods (SN [45], GConv [48], EC [27], RFR [19], HR [38], CTSDG [9], MAT [20]) in terms of number of trainable parameters (*x-axis*), number of operations (GMAC) (*y-axis*) and run time complexity in *seconds per image* (*bubble size*).

The evolution of the attention mechanism (Transformers) [37] helps the tasks, where attention plays a key role, to deal with the non-local modelling. It has turned the image reconstruction task into a different realm. In this context, the transformer-based approaches [56, 50] are proposed for image inpainting. These approaches simply use the commonly utilized transformer block (LN→MSA→LN→FFN) repeatedly in turn increasing the computational cost of the

overall architecture. Also, the quadratic computational complexity of the transformers limits these methods to apply attention to the deep feature maps only with small scales. So the generated image lacks detailed information. To overcome this, Li *et al.* [20] proposed a vanilla transformer block along with the multi-head contextual attention. This method provides better results as compared to existing methods but lacks semantic context understanding. Also, when the transformer block is considered for processing the input with a large size, there may be a chance of information loss due to the quantization of input to a smaller size. This phenomenon is taken care of in [24], by considering a patch-based auto-encoder.

The attention mechanism in general is an effective approach for image inpainting task. Since it helps the network to effectively extract the features from valid locations for inpainting the holes. Also, to fill the holes of large size, it is necessary to have a varying receptive field in consideration while extracting the features. The multi-head attention urges to weigh the feature maps with the valid features. Considering these points, in this work, we propose a nested deformable multi-head attention layer (NDMAL) to transfer the encoder features for effective reconstruction while considering diverse receptive fields. Inspired by the success of linear unified nested attention [26] for a sequence modelling task, we propose a nested deformable multi-head attention layer for image inpainting task. • Unlike [26], we consider encoder and decoder features as packed and unpacked inputs. Though, encoder and decoder features are inputs to the multi-head attention, our proposed layer has linear complexity. Since we utilize the channel-wise attention instead of spatial attention. The proposed NDMAL helps the network to effectively extract the features from the valid region (background) to fill the holes. • Further, we propose deformable multi-head attention (DMHA) for extracting decoder features from diverse fields which are then merged with the skip features from the encoder. Also, a gated feed-forward layer is utilized to again pass the weighted features for reconstruction. Resembling the encoder skip features as a query sequence, packed attention is calculated, called packed context. This packed context is again processed through DMHA with query sequence as decoder features and generated an unpacked context. Both of these packed and unpacked contexts are merged and then forwarded to the next layer. These packed and unpacked context features assist in the effective reconstruction of the inpainted image. The main contributions of our work are:

- Formulating a lightweight architecture consisting of novel transformer layer for facial image inpainting.
- We propose a nested deformable multi-head attention transformer layer (NDMAL) to effectively fuse the encoder and decoder features. The use of NDMAL allows the network to effectively capture long term dependencies and to extract the valid features from maximum receptive fields.
- We propose the analysis of inpainting methods on seen and unseen types of masks.

The ablation study is carried out to verify the efficiency of the proposed NDMAL. Comparative analysis of the proposed approach on Celeb_HQ dataset corrupted with masks from two different datasets and Places2 dataset proves its efficiency for image inpainting task.

2. Related Work

Image inpainting is a sempiternal problem of image restoration where the image with holes is filled with the most relevant content. Earlier works used patch [5], exemplar [17] and diffusion [1] based approaches to inpaint the image. These approaches mainly use the textural or structural statistics of the patches or valid regions to inpaint the hole region. Jin *et al.* [14] proposed a patch-sparsity based method with deduced directional derivatives for image inpainting. Barnes *et al.* [2] proposed a patch-based method where the patch from nearest neighbour match is used for inpainting the image. Though these conventional methods reproduce the inpainted image, they lack the structural consistencies in the outputs.

The deep learning approaches for image inpainting come up with visually plausible results and encounter the traditional inpainting methods. The adversarial training approach provides plausible outcomes in the image-to-image translation task [8, 29, 16]. The very first adversarial training-based approach was proposed by Pathak *et al.* [28] for image inpainting. Later, various methods were proposed with local-global discrimination approach [12], partial convolutions [22], gated convolutions [48], contextual attentions [23, 44, 46, 47] *etc.* for image inpainting. Also, some prior information-based methods were proposed for image inpainting with structurally plausible outcomes [27, 33]. In line to this, the progressive [18] and recurrent [19] approaches were proposed for image inpainting. The parallel processing of multi-resolution features was performed in [41] for robust semantic and plausible texture generation. In [52], the authors proposed a patch-borrowing mechanism for an attention-free generator network with a supplementary reconstruction task that performs as training loss for inpainting. In similar way, the self distillation approach was proposed by Suin *et al.* [35] for image inpainting. The separate generation of textural and structural information for inpainting an image is carried out with separate networks in [9]. With the advantage of transforms to synthesize the features, Yu *et al.* [49] and Suvorov *et al.* [36] proposed the wavelet features and Fast Fourier Convolution based methods respectively to inpaint the images with large masks.

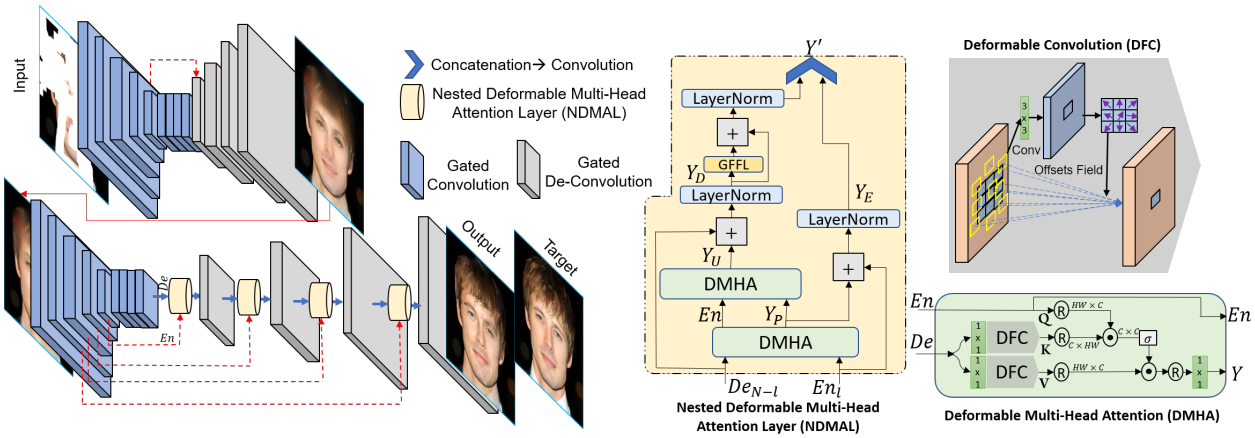

Figure 2. Proposed architecture for image inpainting. We propose a nested deformable multi-head attention transformer layer (NDMAL) to focus on large receptive fields with long term dependencies. The proposed layer consists of single layer in turn reducing the computational complexity of the network.

With the exceptional ability to model the long-term relationship, the transformers are in high demand for various vision applications [3, 7, 40]. Recently, Wan *et al.* [39] proposed a pluralistic image completion with the help of bidirectional attention. Further, Yu *et al.* [50] proposed auto-regressive transformer based pluralistic image inpainting. Similarly, Dong *et al.* [6] proposed an incremental transformer-based three-stage architecture with structure and texture restoration by a transformer and FFT CNN respectively. Zhao *et al.* [54] proposed cross semantic attention layer for diverse image inpainting. The authors in [20] proposed a mask-aware transformer in which the attention module fuses the information from the partial valid tokens. In this work, we propose a modified attention layer named nested deformable multi-head attention layer (NDMAL) to process the encoder and decoder features with nested attention. This layer helps to extract valid attention from diverse receptive fields to inpaint the images faithfully. The detailed exposition of the proposed method is given in Section §3.

3. Proposed Method

In this section we first introduce in general multi-head attention used in the transformer [37], the linear unified nested attention [26] and then we put a light on the proposed nested deformable multi-head attention layer (NDMAL) used for image inpainting task.

3.1. Transformer with Self Attention

The multi-head attention [37] maps $A \in \mathbb{R}^{n\times p} \times B \in \mathbb{R}^{m\times p} \rightarrow Y \in \mathbb{R}^{n\times p}$ is generally formulated as:

$$Y = Attn(A,B) = \sigma(\frac{A\phi_q(B\phi_k)^T}{\sqrt{d_k}})B\phi_v \qquad (1)$$

where, A and B are the query and context sequences with length n and m respectively, σ is the softmax activation, p is the embedding dimension, ϕ_q, ϕ_k and ϕ_v are the trainable parameters used to project the input into query, key and values, d_k is dimension of key. In [37] for multi-head attention $A = B$ is considered, called as *self-attention*. The output of this multi-head attention *i.e.,*self-attention is fed to position wise feed-forward layer followed by layer normalization. The final output of the transformer (Y') is given as:

$$Y' = \eta(FFN(Y_A) + Y_A) \qquad (2)$$

where, η is `LayerNormalization`, $Y_A = \eta(Y + A)$. These transformer layers are sequentially utilized l times in each block. The feed-forward network (FFN) is independently applied on each position and layer normalization controls the gradient scales [37]. The SA generally has quadratic complexity. The computational load of the SA is reduced with applying the SA on small spatial *window size*, $ws = 8 \times 8$ [21, 43] instead of global attention.

3.2. Linear Unified Nested Attention

The linear unified nested attention [26] (LUNA) deals with the quadratic memory and computational complexity of transformers ($\mathcal{O}(mn)$) (§3.1) by introducing an extra input sequence of fixed length by generating two outputs. This in turn gives linear complexity to the transformer layer. The pack (Y_P) and unpack (Y_U) attentions are introduced as:

$$Y_P = Attn(C,B); \quad Y_U = Attn(A,Y_P) \qquad (3)$$

where, $C \in \mathbb{R}^{l\times p}$ is an extra input sequence with fixed length l. The packed and unpacked attentions have the complexity of $\mathcal{O}(lm)$ and $\mathcal{O}(ln)$. So, the LUNA takes three inputs in general (A, B and C) and produces a packed and unpacked attention as output. The LUNA layers take these attentions to further process via FFN and

LayerNormalization as:

$$\begin{aligned} Y_P, Y_U &= LunaAttn(A, C, B) \\ Y_A, C_A &= \eta(Y_P + A), \eta(Y_U + C) \\ Y', C' &= \eta(FFN(Y_A) + Y_A), C_A \end{aligned} \tag{4}$$

where, Y' and C' are the outputs of the LUNA Layer.

3.3. Proposed Nested Deformable Multi-head Attention

In combination to both of the multi-head attention (§3.1) and LUNA attention (§3.2), we propose a nested deformable multi-head attention for the task of image inpainting (Figure 2). The LUNA attention provides an extra input with actual inputs to have linear complexity. Applying the self-attention to the image inpainting task may provide relative contextual information either from the encoded features or from decoder features. Whereas, in our proposed approach, we provide the decoder (De) and skip connection features from the encoder (En) as input. Considering both the features from the encoder and decoder may allow delving into the valid feature space efficiently. Also, in order to extract maximum receptive field from the decoder processed features, we leverage the deformable convolution layer [4] unlike [37] and [26]. Here, we consider the encoder features to be the context information provided to the decoder for effective reconstruction. So, the proposed deformable multi-head attention (DMHA) is formulated as:

$$\begin{aligned} Y = DMHA(De_{N-l}, En_l) = \\ \sigma\left(En_l \phi_q (De_{N-l} \phi_k^{df})^T\right) De_{N-l} \phi_v^{df} \end{aligned} \tag{5}$$

where, ϕ^{df} shows the deformable convolution applied to the decoder features to delve into the maximum receptive fields, $l \in (1, 4)$ is the number of layers and $N = 5$ (*see DMHA in Figure 2*). In deformable convolution, the normal grid $O = \{(-1,-1),(-1,0),...,(0,1),(1,1)\}$ is augmented with the offsets $\{\triangle p_n | n = 1, ..., P\}$, $P = |O|$. So, for each location p_0 in the output feature map ϕ^{df},

$$\phi^{df}(p_0) = \sum_{p_n \in O} w(p_n).x(p_0 + p_n + \triangle p_n) \tag{6}$$

Further, we introduce the nested deformable attention mechanism to increase the required receptive field and to focus on long-term dependencies. Also, nesting of DMHA makes sense that, it can capture sufficient contextual information. The packed (Y_P) and unpacked (Y_U) outcomes of the nested deformable attention are given as:

$$\begin{aligned} Y_P &= DMHA(De_{N-l}, En_l) \\ Y_U &= DMHA(Y_P, De_{N-l}) \end{aligned} \tag{7}$$

Since we consider the encoder layer features with the input sequence, it will be able to pack the global context of the input efficiently. The packed and unpacked outputs are then forwarded to layer normalization and gated feed-forward layer (GFFL). The output (Y') of proposed NDMAL is given as:

$$\begin{aligned} Y_E, Y_D &= \eta(Y_P + En_l), \eta(Y_U + De_{N-l}) \\ Y' &=< \eta(GFFL(Y_D) + Y_D), Y_E > \end{aligned} \tag{8}$$

where, $< . >$ indicates concatenation operation. The GFFL is the gated feed forward layer which is used to suppress any undesired features if present. The GFFL is represented as:

$$GFFL(f_{in}) = \phi(f_{in}) + \mathbb{G}(\psi(f_{in})) \tag{9}$$

where, $\mathbb{G}$ is GELU activation function, ϕ and ψ are learnable parameters.

3.4. Overall Architecture

The overall architecture of the proposed approach is visualized in Figure 2. We follow a coarse-to-fine architecture. The purpose behind the coarse-to-fine architecture is to forward the coarse output features through the proposed NDMAL as a query to provide sufficient contextual information. So that the network will be able to capture long-term dependencies effectively. The proposed NDMAL is utilized in the fine stage which takes input from the encoder layer and considers it as a query to the respective decoder feature key and values. Also, the packed attention in the NDMAL is calculated with respect to the encoder skip inputs which is then concatenated with the processed unpacked attention. The concatenation of both allows to preserve the valid content efficiently.

The encoder and decoder layers of both the coarse and fine stages are designed with the gated convolution layer followed by a LeakyReLu activation. The successive encoder layers at the bottleneck of the coarse stage allow focusing on the different receptive fields which produce an approximate output. This coarse output is then fed to the fine stage which includes the proposed NDMAL. The overall architecture with effective usage of NDMAL generates faithful inpainted results. As we are considering the deformable multi-head attention, it may help the network to extract information from maximum receptive fields. Also, the nested multi-head attention applied to the encoded and decoded features may help to capture the long-term dependencies. So, unlike existing transformer architectures, our proposed NDMAL consists of only one block with $ws = 8 \times 8$. This helps to reduce the computational cost of our proposed inpainting network. Though the two inputs to the proposed NDMAL are having the length of n, m, it preserves the linear complexity. This is because we apply the attention channel-wise instead of spatially [51]. So, the attention will effectively encode the global context by computing the cross-covariance across the channels. This also reduces the necessity of an extra input with constant length (l) like [26].

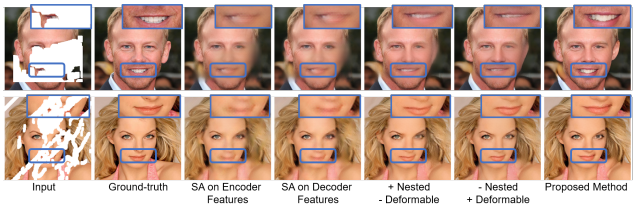

Figure 3. Analysis on different configurations of proposed method (*Note: + indicates inclusion and - indicates exclusion of particular block, SA is self-attention*).

Configuration (Parameters)	PSNR	SSIM	L1	LPIPS	FID
SA on En Feat (3.61M)	24.25	0.842	4.259	0.162	9.482
SA on De Feat (3.61M)	24.98	0.857	4.008	0.151	9.106
+Nested -Deformable (3.62M)	27.68	0.915	3.007	0.104	7.864
-Nested +Deformable (3.85M)	26.28	0.897	3.856	0.122	8.567
Proposed Network (4.12M)	**28.19**	**0.931**	**2.575**	**0.082**	**6.844**

Table 1. Quantitative comparison for different configurations of the proposed network for image inpainting on $0.01 - 0.6$ mask ratio on CelebA-HQ dataset (*Note: + indicates inclusion and - indicates exclusion of particular block, SA is self-attention, En Feat and De Feat are encoder and decoder features respectively*).

4. Training of the Proposed Network

The proposed architecture is trained with the corrupted image and its mask as input and generates an inpainted image as output. The discriminator network is the same as that of [13]. While training, the image values are linearly scaled between the range $[0:1]$. Weight parameters of the network are updated on NVIDIA DGX station having Tesla V100 1×16 GB GPU with the batch size of 1 for 200 epochs (38 GPU Hours). The ADAM optimizer [15] with the learning rate of 2×10^{-4}, $\beta_1 = 0.5$ and $\beta_2 = 0.99$ is used.

4.1. Loss Functions

Given the corrupted image with holes (I_C) and the mask (I_M) with ones at holes and zeros at the non-hole region, it is required to generate the inpainted image (I_I) similar to the target image (I_T). The L_1 loss is used to optimize the network for better reconstruction. For the generation of the globally and locally consistent realistic image, the adversarial loss plays an important role [8], [13]. The adversarial loss is the min-max problem between generator and discriminator, respectively and given as:

$$L_{GAN} = \max_{\mathbf{D}} \min_{\mathbf{G}} \mathbb{E}[log(D(I_C, I_T))] + \mathbb{E}[log(1 - D(I_C, G(I_C)))] \quad (10)$$

where, D is the discriminator and G is the generator. To guide the network for textural and structural information, the perceptual loss is calculated between the deep feature maps of the ground-truth and inpainted images by passing them through the pre-trained VGG19 model [34] as:

$$L_P = \sum_{s=1}^{S} (\|\phi_s(I_T) - \phi_s(I_I)\|_1) \quad (11)$$

where, ϕ_s are the feature maps ($s \in (1, S)$) of the VGG19 model. The edge loss is also considered to focus on the edge enhancement while training. The edge loss with sobel operator $\mathbb{S}$ is formulated as:

$$L_e = \|\mathbb{S}(I_T) - \mathbb{S}(I_I)\|_1 \quad (12)$$

So, the overall loss for training the network is given as:

$$L_{Total} = \lambda_1 L_1 + \lambda_{GAN} L_{GAN} + \lambda_e L_e + \lambda_P L_P \quad (13)$$

here, λ_{loss} are the weights assigned for the respective loss functions. The values (*determined experimentally*) of each of the weights are $\lambda_1 = 10$, $\lambda_e = 2$, $\lambda_P = 3$, and $\lambda_{GAN} = 0.1$ (*analysis of effect of each loss function is provided in supplementary material*).

5. Experiments

Here, we provide details of datasets and metrics used to compare proposed approach with baselines, the ablation study on different configurations of proposed architecture, comparative and computational complexity analysis.

5.1. Datasets, Metrics and Baselines

This work focuses on the facial image inpainting. For this purpose, we use a publicly available celebrity faces dataset named CelebA-HQ [25]. This dataset consists of 28k images for training and 2k images for testing. A natural image dataset *i.e.,* Places2 [57] which contains images from 365 different places is also used. To corrupt the face images, we used two different types of the mask datasets. The NVIDIA mask dataset [22] and quick draw irregular mask dataset (QD-IMD) [10]. The natural images are corrupted using NVIDIA masks. The testing set of NVIDIA mask dataset covers different hole-to-image area *i.e.,* mask ratios in the range $(0.01, 0.6]$. In total, there are 12k masks available which are divided into six sets with $(0.01, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.4]$, $(0.4, 0.5]$, and $(0.5, 0.6]$ mask ratio. Also, a mask dataset with strokes drawn by human hand called as quick draw irregular mask dataset (QD-IMD) [10] is used for the evaluation of the proposed architecture. The two mask datasets differ from each other where, the NVIDIA mask dataset is based on occlusion/dis-occlusion mask estimation between two consecutive frames which has sharp edges due of rough crops near to borders and the QD-IMD consists of irregularly drawn strokes without sharp edges. *The sample masks of both the datasets are given in supplementary material.*

For quantitative evaluation, we consider five evaluation measures: (i) peak-signal-to-noise ratio (PSNR), (ii) structural similarity index (SSIM), (iii) L_1 norm, (iv) Perceptual

Mask Ratio	Metric	SN [45] ECCV-18	GMCNN [42] NIPS-18	PIC [55] CVPR-19	Gconv [48] ICCV-19	EC [27] CVPRW-19	RFR [19] CVPR-20	HR [38] WACV-21	CTSDG [9] ICCV-21	MAT [20] CVPR-22	Ours
0.01-0.2	PSNR↑	30.84	30.54	32.08	32.06	32.04	33.45	33.28	33.57	33.56	33.99
	SSIM↑	0.961	0.957	0.967	0.960	0.973	0.973	0.976	0.979	0.977	0.982
	L_1 ↓	2.827	2.867	2.689	2.681	3.108	1.824	1.925	1.329	1.147	1.017
	LPIPS↓	0.060	0.057	0.043	0.039	0.038	0.045	0.041	0.029	0.027	0.022
	FID↓	4.134	7.537	4.042	4.309	4.042	2.516	2.257	2.105	2.032	1.775
0.2-0.4	PSNR↑	25.77	24.49	25.30	25.48	26.30	26.44	26.76	27.02	27.13	27.43
	SSIM↑	0.896	0.894	0.891	0.904	0.901	0.917	0.935	0.936	0.931	0.948
	L_1 ↓	4.246	4.120	3.691	4.147	3.194	3.022	3.213	2.466	2.466	2.382
	LPIPS↓	0.2091	0.1711	0.1772	0.1668	0.1630	0.1414	0.1341	0.1020	0.0944	0.0740
	FID↓	10.643	28.170	14.376	11.010	7.338	11.767	10.330	7.516	6.620	5.862
0.4-0.6	PSNR↑	18.65	18.74	19.01	19.70	21.33	21.23	22.04	22.24	22.55	23.14
	SSIM↑	0.657	0.744	0.679	0.840	0.809	0.755	0.831	0.845	0.847	0.858
	L_1 ↓	8.852	6.7465	7.0105	5.6945	5.828	6.354	5.3445	4.451	4.4015	4.326
	LPIPS↓	0.3690	0.4060	0.3451	0.3017	0.2755	0.2551	0.2429	0.1910	0.1811	0.1479
	FID↓	61.160	50.981	49.120	34.940	33.011	30.650	28.498	14.371	13.121	12.897

Table 2. Quantitative comparison of the proposed method (Ours) with the state-of-the-art methods on NVIDIA [22] masks for image inpainting on CelebA-HQ dataset (↑- *Higher is better,* ↓- *Lower is better*). The best and second best results are in red and blue.

Figure 4. Qualitative comparison of the proposed method (Ours) with existing methods on CelebA_HQ dataset for NVIDIA [22] mask.

Image Patch Similarity (LPIPS) [53] to analyse the perceptual similarity between inpainted and ground-truth images, and (v) Fréchet inception distance (FID) [11] to quantify the distance between distributions of inpainted and ground-truth images.

To examine the efficiency, we consider the comparison of our proposed method with existing state-of-the-art methods for image inpainting : Shift-net (SN) [45], GMCNN:NIPS-18 [42], pluristic-image completion (PIC) [55], gated-convolutions (Gconv) [48], edge-connect (EC) [27], recurrent feature reasoning (RFR) [19], hypergrphs (HR) [38], contextual texture-structure dual generation (CTSDG) [9], and mask aware transformers (MAT) [20].

5.2. Ablation Study

In order to come up with an optimum architecture for image inpainting task, we carried out meticulous experiments with different combinations of our network. These experiments include, (a) considering the self attention (§3.4) applied on the encoder features and merged with decoder features (*SA on encoder features*) , (b) self attention (§3.4) applied on the decoder features and merged with encoder features (*SA on decoder features*), (c) applying the nested attention without deformable layer (similar to LUNA §3.2) (*+Nested -Deformable*), (d) applying the deformable multi-head attention without nested attention layers (*-Nested +Deformable*), (e) finally, applying the nested deformable multi-head attention layer (*+Nested +Deformable i.e., Proposed Network*) (*see Table 1*). *The architectural differences of blocks used for ablation experiments are given in supplementary material.*

Purpose of this study is to compare quantitative and qualitative differences between different configurations of the proposed network. **We examine whether the self attention applied on either encoder or decoder features works better.** The existing self attention tries to extract the long term dependencies from the input feature maps. Applying it on the encoder or decoder features affects differently while reconstructing the image. Row 2 and 3 in Table 1 show the results for the configuration where the self attention is applied

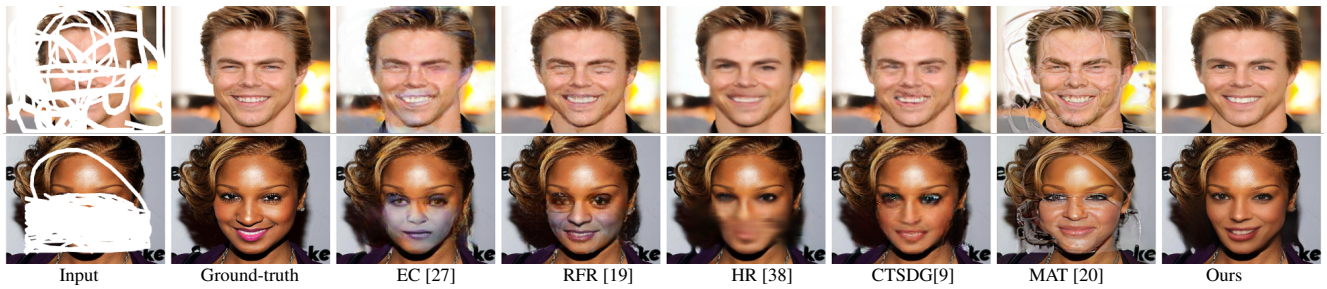

Figure 5. Qualitative comparison of the proposed method (Ours) with existing methods on CelebA_HQ dataset for unknown mask dataset QD-IMD [10].

Mask Ratio	Method	PSNR↑	SSIM↑	L_1 ↓	LPIPS↓	FID↓
0.01-0.2	EC [27]	33.19	0.972	1.340	0.0404	2.929
	RFR [19]	33.45	0.973	1.824	0.0291	2.516
	HR [38]	33.28	0.974	1.143	0.0259	2.051
	CTSDG[9]	34.55	0.981	0.984	0.0186	1.913
	MAT [20]	34.66	0.982	0.945	0.0201	1.627
	Ours	35.05	0.989	0.818	0.0172	1.567
0.2-0.4	EC [27]	25.85	0.933	2.719	0.1319	7.561
	RFR [19]	26.92	0.939	2.513	0.1182	7.267
	HR [38]	27.68	0.948	2.443	0.1082	6.652
	CTSDG[9]	28.48	0.956	2.089	0.0540	6.262
	MAT [20]	28.62	0.957	1.930	0.0535	6.016
	Ours	28.94	0.961	1.807	0.0533	5.181
0.4-0.6	EC [27]	22.43	0.856	5.007	0.2136	19.543
	RFR [19]	22.93	0.868	4.754	0.1801	18.650
	HR [38]	23.37	0.871	4.039	0.1734	17.685
	CTSDG[9]	23.80	0.880	3.707	0.1308	16.111
	MAT [20]	24.03	0.887	3.637	0.1229	15.921
	Ours	24.56	0.895	3.508	0.1186	15.493

Table 3. Quantitative comparison of the proposed method (Ours) with state-of-the-art methods on QD-IMD [10] masks for image inpainting on CelebA-HQ dataset.

on encoder and decoder features respectively. From Table 1 and Figure 3, it is clear that, the self attention when applied with encoder (*row 2 of Table 1*) or decoder (*row 3 of Table 1*) feature map as input fails to produce efficient outcome in terms numeric and visual results. Inspired with the LUNA attention, **we ought to include the LUNA layer in the inapinting architecture to verify its ability to delve into the valid features**. Contrary to self attention, the results are improved quantitatively and also generate better structural information visually (*see row 4 in Table 1 and column 5 in Figure 3*). The reason behind this might be, here we consider both the information from encoder features and decoder features in order to get better contextual information as compared to considering either of them. Further, we pondered that, if we try to consider maximum receptive field, it will further help the network towards better outcome. **A study is carried out to determine whether addition of deformable convolution works well to extract maximum receptive field.** In light of that, we considered a deformable multi-head attention (*row 5 of Table 1*) for extracting the contextual information from the input feature maps which resulted into better convergence of structural information. So, in combination to *+Nested* and *+Deformable* (*see Proposed Network in Table 1 and Figure 3*), we come up with our proposed network, nested deformable multi-head attention layer (NDMAL) for image inpainting. This proposed NDMAL gives inpainted output akin to ground-truth.

5.3. Comparative Analysis

We train our network on CelebA-HQ image dataset corrupted with NVIDIA mask training dataset similar to baselines (§5.1). For comparative analysis, we considered two types of masks as mentioned in §5.1. For both mask datasets, we considered $0.01 - 0.2$, $0.2 - 0.4$ and $0.4 - 0.6$ mask ratios. Quantitative comparison of the proposed method with existing baselines in terms of PSNR, SSIM, L_1 norm, LPIPS and FID is given in Table 2. From Table 2, we can clearly mention that the proposed method effectively outperforms all the baselines for all mask ratios and ultimately on average of all the mask ratios. Along with the numeric superiority, we assess visual comparison of proposed method with existing baselines. Visual comparison is depicted in Figure 4. With the comparison, we come up with some observations: our proposed method does not generate ghosting outcomes, it does not create stitching effects, it does not produce over sharp results, *etc.* Furthermore, our outputs are more accurate when compared with baselines because their resemblance to ground truth is greater.

Along with this comparison on NVIDIA dataset masks, we urge to verify reliability of our method with other mask datasets. For this experiment, we consider the CelebA-HQ images corrupted with QD-IMD dataset. ***Similar to existing baselines, our model is also not trained for these type of masks.*** It means, we are comparing all the methods (including ours) with unknown types of masks. *In order to make it simple, we compare our method with only best five baselines.* The quantitative and qualitative results' comparison is provided in Table 3 and Figure 5 respectively. Our proposed approach gives quantitatively improved results as compared with the existing baselines. In Figure 5 we can see that, comparing our results with existing best methods, we find that ours go more in the direction of plausible gener-

Mask Ratio	Metric	SN [45]	GMCNN [42]	PIC [55]	Gconv [48]	EC [27]	RFR [19]	HR [38]	CTSDG [9]	MAT [20]	Ours
0.01-0.2	PSNR	27.88	28.22	29.52	29.50	29.69	30.64	30.12	30.61	31.68	32.51
	SSIM	0.876	0.894	0.917	0.921	0.915	0.928	0.936	0.953	0.954	0.968
	L1	3.371	3.637	2.796	2.698	2.585	1.181	1.661	1.490	1.121	1.104
	LPIPS	0.134	0.113	0.136	0.127	0.132	0.102	0.098	0.066	0.044	0.062
	FID	10.763	10.543	8.447	7.718	7.499	6.104	6.148	4.459	3.696	3.639
0.2-0.4	PSNR	22.67	22.82	23.46	22.80	23.70	24.22	24.18	25.10	25.73	26.22
	SSIM	0.816	0.858	0.842	0.872	0.877	0.850	0.856	0.877	0.884	0.893
	L1	5.173	5.532	4.410	4.393	4.081	3.828	3.638	3.327	3.067	2.661
	LPIPS	0.239	0.223	0.218	0.205	0.203	0.193	0.184	0.183	0.166	0.174
	FID	29.126	27.398	25.799	22.007	21.018	20.218	19.326	18.427	14.839	14.254
0.4-0.6	PSNR	18.19	18.19	18.82	19.48	19.52	20.76	20.83	21.03	21.18	21.89
	SSIM	0.621	0.660	0.692	0.724	0.719	0.726	0.745	0.770	0.726	0.776
	L1	9.33	7.4985	7.111	6.6565	6.361	6.486	5.999	5.7625	5.333	5.037
	LPIPS	0.447	0.400	0.371	0.357	0.360	0.343	0.335	0.329	0.248	0.312
	FID	74.150	73.696	73.408	68.005	54.341	49.204	55.461	40.266	35.810	37.887

Table 4. Quantitative comparison of the proposed method (Ours) with the state-of-the-art methods on NVIDIA [22] masks for image inpainting on Places2 dataset.

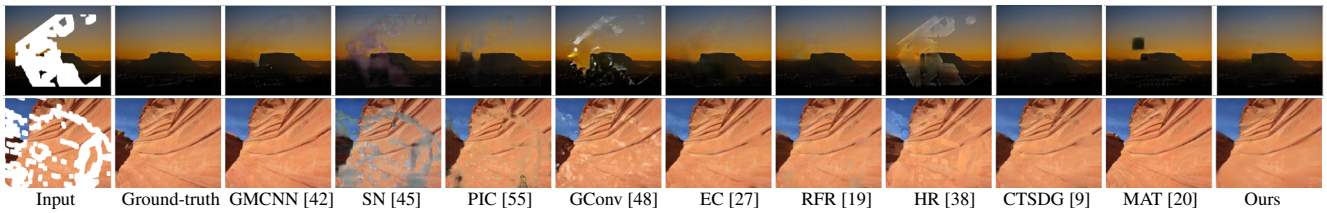

Figure 6. Qualitative comparison of the proposed method (Ours) with existing methods on Places2 dataset for NVIDIA [22] mask.

ation. We, give this credit of faithful image inpainting to out proposed nested deformable multi-head attention. Since, it is able to easily extract the contextual information from both the encoded features and decoded features.

To show the generalizability of our proposed method, we have considered a Places2 natural images dataset [57]. The quantitative and qualitative comparison on Places2 dataset is given in Table 4 and Figure 6 respectively. This comparison shows that our proposed method performs well for the non-face/natural image inpainting. *Though our proposed method has very less number of parameters (4.1M) i.e.,* $\frac{1}{15}^{th}$ *of the baseline [20] (60M), it performs well for face and non-face image inpainting task.*

5.4. Complexity Analysis

We claim that, our propose method has low complexity with good results as compared to existing baselines. Our proposed nested deformable multi-head attention has linear complexity, since we apply the attention across channels similar to [51]. Also, the existing self attention based methods utilize number of blocks with different window sizes to capture long term dependencies in turn increasing the computational cost. Here, in this approach we come up with a single block in our NDMAL as it already consider two different feature maps to find the relative contextual information. Further, the nested attention helps the layer to extract valid content more extensively. Also, the deformable additively provide it with the larger receptive field. These points altogether allow a single block NDMAL with a $ws = 8$ to extract relevant features for image inpainting.

The computational complexity analysis in terms of number of trainable parameters, number of operations *i.e.,* Giga multiply-accumulate operations (GMAC) and average run time in terms of seconds/image is visualized in Figure 1. From Figure 1 and Tables 2, 3 and 4, it is clear that, with lower computational complexity, our method has good performance as compared to existing baselines (*the detailed quantitative values of Figure 1 are provided in supplementary material*).

6. Conclusion

This work aimed to propose a lightweight architecture with a novel transformer layer for facial image inpainting. To do this, we proposed a nested deformable multi-head attention layer with a capacity of extracting valid features from maximum receptive fields and capturing long term dependencies effectively. The proposed method is compared quantitatively and qualitatively with existing state-of-the-art methods for image inpainting on CelebA_HQ and Places2 dataset corrupted using NVIDIA mask dataset. To verify the reliability, we compared the proposed method with existing methods on CelebA_HQ corrupted using unknown masks from QD-IMD dataset. The experiments show the effectiveness of proposed method for facial and non-facial image inpainting.

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] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

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

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE ICCV*, pages 764–773, 2017.

[5] 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.

[6] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE CVPR*, pages 11358–11368, 2022.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE CVPR*, pages 12873–12883, 2021.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[9] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE ICCV*, pages 14134–14143, 2021.

[10] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017.

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

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

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

[14] Darui Jin and Xiangzhi Bai. Patch-sparsity-based image inpainting through a facet deduced directional derivative. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(5):1310–1324, 2018.

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

[16] Ashutosh Kulkarni, Prashant W. Patil, and Subrahmanyam Murala. Progressive subtractive recurrent lightweight network for video deraining. *IEEE Signal Processing Letters*, 29:229–233, 2022.

[17] 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.

[18] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE ICCV*, pages 5962–5971, 2019.

[19] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE CVPR*, pages 7760–7768, 2020.

[20] 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 CVPR*, pages 10758–10768, 2022.

[21] 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 ICCV*, pages 1833–1844, 2021.

[22] 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 ECCV*, pages 85–100, 2018.

[23] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE ICCV*, pages 4170–4179, 2019.

[24] 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 CVPR*, pages 11347–11357, 2022.

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE ICCV*, pages 3730–3738, 2015.

[26] Xuezhe Ma, Xiang Kong, Sinong Wang, Chunting Zhou, Jonathan May, Hao Ma, and Luke Zettlemoyer. Luna: Linear unified nested attention. *Advances in Neural Information Processing Systems*, 34:2441–2453, 2021.

[27] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE ICCV Workshops*, pages 3265–3274, 2019.

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE CVPR*, pages 2536–2544, 2016.

[29] Prashant Patil and Subrahmanyam Murala. Fggan: A cascaded unpaired learning for background estimation and foreground segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1770–1778, 2019.

[30] Shruti S Phutke and Subrahmanyam Murala. Diverse receptive field based adversarial concurrent encoder network for

image inpainting. *IEEE Signal Processing Letters*, 28:1873–1877, 2021.

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

[32] Shruti S Phutke and Subrahmanyam Murala. Image inpainting via spatial projections. *Pattern Recognition*, 133:109040, 2023.

[33] 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 ICCV*, pages 181–190, 2019.

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

[35] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Distillation-guided image inpainting. In *Proceedings of the IEEE ICCV*, pages 2481–2490, 2021.

[36] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *2022 IEEE WACV*, pages 3172–3182. IEEE, 2022.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[38] Gourav Wadhwa, Abhinav Dhall, Subrahmanyam Murala, and Usman Tariq. Hyperrealistic image inpainting with hypergraphs. In *Proceedings of the IEEE WACV*, pages 3912–3921, 2021.

[39] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE ICCV*, pages 4692–4701, 2021.

[40] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE CVPR*, pages 1571–1580, 2021.

[41] Wentao Wang, Jianfu Zhang, Li Niu, Haoyu Ling, Xue Yang, and Liqing Zhang. Parallel multi-resolution fusion network for image inpainting. In *Proceedings of the IEEE ICCV*, pages 14559–14568, 2021.

[42] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. *arXiv preprint arXiv:1810.08771*, 2018.

[43] 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 CVPR*, pages 17683–17693, 2022.

[44] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *Proceedings of the IEEE ICCV*, pages 8858–8867, 2019.

[45] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *Proceedings of the ECCV*, pages 1–17, 2018.

[46] 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 CVPR*, pages 7508–7517, 2020.

[47] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE CVPR*, pages 5505–5514, 2018.

[48] 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 ICCV*, pages 4471–4480, 2019.

[49] Yingchen Yu, Fangneng Zhan, Shijian Lu, Jianxiong Pan, Feiying Ma, Xuansong Xie, and Chunyan Miao. Wavefill: A wavelet-based generation network for image inpainting. In *Proceedings of the IEEE ICCV*, pages 14114–14123, 2021.

[50] 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.

[51] 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 CVPR*, pages 5728–5739, 2022.

[52] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *Proceedings of the IEEE ICCV*, pages 14164–14173, 2021.

[53] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE CVPR*, pages 586–595, 2018.

[54] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE CVPR*, pages 5741–5750, 2020.

[55] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE CVPR*, pages 1438–1447, 2019.

[56] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Tfill: Image completion via a transformer-based architecture. *arXiv preprint arXiv:2104.00845*, 2021.

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