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 CelebHQ [25] dataset using known (NVIDIA) and unknown (QD-IMD) masks and Places2 [57] dataset with NVIDIA masks along with an 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.

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. Reassembling 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 semiprincipal 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:

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

where, $A$ and $B$ are the query and context sequences with length $n$ and $m$ respectively, $\sigma$ is the softmax activation, $p$ is the embedding dimension, $\phi_q$, $\phi_k$ and $\phi_v$ 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:

$$Y' = \eta(\text{FFN}(Y_A) + Y_A)$$

where, $\eta$ 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 ($O(mn)$) (§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:

$$Y_P = \text{Attn}(C, B); \quad Y_U = \text{Attn}(A, Y_P)$$

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 $O(lm)$ and $O(ln)$. 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
LayerNorm as:

\[ Y_P, Y_U = LunaAttn(A, C, B) \]
\[ Y_A, C_A = \eta(Y_P + A), \eta(Y_U + C) \]
\[ Y', C' = \eta(FFN(Y_A) + Y_A), C_A \]

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:

\[ Y = DMHA(De_{N-1}, En_l) = \sigma \left( En_l \phi_y(De_{N-l} \phi_{df}^y)^T \right) De_{N-l} \phi_{df}^y \]  

(5)

where, \( \phi_{df} \) 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 \( O = \{(-1, -1), (-1, 0), \ldots, (0, 1), (1, 1)\} \) is augmented with the offsets \{\( \Delta p_n | n = 1, \ldots, P \}\), \( P = |O| \). So, for each location \( p_0 \) in the output feature map \( \phi_{df} \),

\[ \phi_{df}(p_0) = \sum_{p_n \in O} w(p_n).x(p_0 + p_n + \Delta p_n) \]  

(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:

\[ Y_P = DMHA(De_{N-1}, En_l) \]
\[ Y_U = DMHA(Y_P, De_{N-l}) \]  

(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:

\[ Y_E, Y_D = \eta(Y_P + En_l), \eta(Y_U + De_{N-l}) \]
\[ Y' = < \eta(GFFL(Y_D) + Y_D), Y_E > \]  

(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:

\[ GFFL(f_{in}) = \phi(f_{in}) + G(\psi(f_{in})) \]  

(9)

where, \( G \) is GELU activation function, \( \phi \) and \( \psi \) 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].
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 $2 \times 10^{-4}$, $\beta_1 = 0.5$ 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 ($I_T$) 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 the generator and discriminator, respectively and given as:

$$L_{GAN} = \max_D \min_G \mathbb{E}[\log(D(I_C, I_T))]$$
$$+\mathbb{E}[\log(1 - D(I_C, G(I_C)))]$$

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:

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

where, $\phi_s$ 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 $S$ is formulated as:

$$L_e = ||S(I_T) - S(I_I)||_1$$

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

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

4.2. Weighting and Mask Ratios

We define five weighting schemes for the proposed loss functions with the weights $\lambda_{GAN} = 0.1$, $\lambda_e = 2$, $\lambda_P = 3$. The NVIDIA mask dataset [22] and quick draw irregular mask dataset (QD-IMD) [10] are used for this purpose. 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.

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).

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

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...
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.

<table>
<thead>
<tr>
<th>Mask Ratio</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>( L_1 )↓</th>
<th>LPIPS↓</th>
<th>FID↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01-0.2</td>
<td>30.84</td>
<td>0.961</td>
<td>2.827</td>
<td>4.134</td>
<td></td>
</tr>
<tr>
<td>0.2-0.4</td>
<td>25.77</td>
<td>0.896</td>
<td>4.246</td>
<td>2.091</td>
<td>10.64</td>
</tr>
<tr>
<td>0.4-0.6</td>
<td>18.65</td>
<td>0.657</td>
<td>8.852</td>
<td>0.3690</td>
<td>61.160</td>
</tr>
</tbody>
</table>

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], picturesque-image completion (PIC) [55], gated-convolutions (Gconv) [48], edge-connect (EC) [27], recurrent feature reasoning (RFR) [19], hypergrahps (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.
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 LAYER in the inpainting 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, L1 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-

<table>
<thead>
<tr>
<th>Mask Ratio</th>
<th>Method</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>L1↓</th>
<th>LPIPS↓</th>
<th>FID↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01-0.2</td>
<td>EC [27]</td>
<td>33.19</td>
<td>0.972</td>
<td>1.340</td>
<td>0.0404</td>
<td>2.929</td>
</tr>
<tr>
<td></td>
<td>RFR [19]</td>
<td>33.45</td>
<td>0.973</td>
<td>1.824</td>
<td>0.0291</td>
<td>2.516</td>
</tr>
<tr>
<td></td>
<td>HR [38]</td>
<td>33.28</td>
<td>0.974</td>
<td>1.143</td>
<td>0.0259</td>
<td>2.051</td>
</tr>
<tr>
<td></td>
<td>CTSGD[9]</td>
<td>34.55</td>
<td>0.981</td>
<td>0.984</td>
<td>0.0186</td>
<td>1.913</td>
</tr>
<tr>
<td></td>
<td>MAT [20]</td>
<td>34.66</td>
<td>0.982</td>
<td>0.945</td>
<td>0.0201</td>
<td>1.627</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>35.05</td>
<td>0.989</td>
<td>0.818</td>
<td>0.0172</td>
<td>1.567</td>
</tr>
<tr>
<td>0.2-0.4</td>
<td>EC [27]</td>
<td>25.85</td>
<td>0.933</td>
<td>2.719</td>
<td>0.1319</td>
<td>7.561</td>
</tr>
<tr>
<td></td>
<td>RFR [19]</td>
<td>26.92</td>
<td>0.939</td>
<td>2.513</td>
<td>0.1182</td>
<td>7.267</td>
</tr>
<tr>
<td></td>
<td>HR [38]</td>
<td>27.68</td>
<td>0.948</td>
<td>2.443</td>
<td>0.1082</td>
<td>6.652</td>
</tr>
<tr>
<td></td>
<td>CTSGD[9]</td>
<td>28.48</td>
<td>0.956</td>
<td>2.089</td>
<td>0.0540</td>
<td>6.262</td>
</tr>
<tr>
<td></td>
<td>MAT [20]</td>
<td>28.62</td>
<td>0.957</td>
<td>1.930</td>
<td>0.0535</td>
<td>6.016</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>28.94</td>
<td>0.961</td>
<td>1.807</td>
<td>0.0533</td>
<td>5.181</td>
</tr>
<tr>
<td>0.4-0.6</td>
<td>EC [27]</td>
<td>22.43</td>
<td>0.856</td>
<td>5.007</td>
<td>0.2136</td>
<td>19.543</td>
</tr>
<tr>
<td></td>
<td>RFR [19]</td>
<td>22.93</td>
<td>0.868</td>
<td>4.754</td>
<td>0.1801</td>
<td>18.650</td>
</tr>
<tr>
<td></td>
<td>HR [38]</td>
<td>23.37</td>
<td>0.871</td>
<td>4.039</td>
<td>0.1734</td>
<td>17.685</td>
</tr>
<tr>
<td></td>
<td>CTSGD[9]</td>
<td>23.80</td>
<td>0.880</td>
<td>3.707</td>
<td>0.1308</td>
<td>16.111</td>
</tr>
<tr>
<td></td>
<td>MAT [20]</td>
<td>24.03</td>
<td>0.887</td>
<td>3.637</td>
<td>0.1229</td>
<td>15.921</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>24.56</td>
<td>0.895</td>
<td>3.508</td>
<td>0.1186</td>
<td>15.493</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01-0.2</td>
<td>PSNR</td>
<td>27.88</td>
<td>28.22</td>
<td>29.52</td>
<td>29.50</td>
<td>29.69</td>
<td>30.64</td>
<td>30.12</td>
<td>30.61</td>
<td>31.68</td>
<td>32.51</td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.876</td>
<td>0.894</td>
<td>0.917</td>
<td>0.921</td>
<td>0.915</td>
<td>0.928</td>
<td>0.936</td>
<td>0.953</td>
<td>0.954</td>
<td>0.968</td>
</tr>
<tr>
<td></td>
<td>L1</td>
<td>3.371</td>
<td>3.637</td>
<td>2.796</td>
<td>2.698</td>
<td>2.585</td>
<td>1.181</td>
<td>1.661</td>
<td>1.490</td>
<td>1.121</td>
<td>1.104</td>
</tr>
<tr>
<td></td>
<td>LPIPS</td>
<td>0.134</td>
<td>0.113</td>
<td>0.136</td>
<td>0.127</td>
<td>0.132</td>
<td>0.102</td>
<td>0.098</td>
<td>0.066</td>
<td>0.044</td>
<td>0.062</td>
</tr>
<tr>
<td>0.2-0.4</td>
<td>PSNR</td>
<td>22.67</td>
<td>22.82</td>
<td>23.46</td>
<td>22.80</td>
<td>23.70</td>
<td>24.22</td>
<td>24.18</td>
<td>25.10</td>
<td>25.73</td>
<td>26.22</td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.816</td>
<td>0.858</td>
<td>0.842</td>
<td>0.872</td>
<td>0.877</td>
<td>0.850</td>
<td>0.856</td>
<td>0.877</td>
<td>0.884</td>
<td>0.893</td>
</tr>
<tr>
<td></td>
<td>LPIPS</td>
<td>0.239</td>
<td>0.223</td>
<td>0.218</td>
<td>0.205</td>
<td>0.203</td>
<td>0.193</td>
<td>0.184</td>
<td>0.183</td>
<td>0.166</td>
<td>0.174</td>
</tr>
<tr>
<td>0.4-0.6</td>
<td>PSNR</td>
<td>18.19</td>
<td>18.19</td>
<td>18.82</td>
<td>19.48</td>
<td>19.52</td>
<td>20.76</td>
<td>20.83</td>
<td>21.03</td>
<td>21.18</td>
<td>21.89</td>
</tr>
<tr>
<td></td>
<td>SSIM</td>
<td>0.621</td>
<td>0.660</td>
<td>0.692</td>
<td>0.724</td>
<td>0.719</td>
<td>0.726</td>
<td>0.745</td>
<td>0.770</td>
<td>0.726</td>
<td>0.776</td>
</tr>
<tr>
<td></td>
<td>L1</td>
<td>9.33</td>
<td>7.4985</td>
<td>7.111</td>
<td>6.6565</td>
<td>6.361</td>
<td>6.486</td>
<td>5.999</td>
<td>5.7625</td>
<td>5.333</td>
<td>5.037</td>
</tr>
<tr>
<td></td>
<td>LPIPS</td>
<td>0.447</td>
<td>0.400</td>
<td>0.371</td>
<td>0.357</td>
<td>0.360</td>
<td>0.343</td>
<td>0.335</td>
<td>0.329</td>
<td>0.248</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>FID</td>
<td>74.150</td>
<td>73.696</td>
<td>73.408</td>
<td>68.005</td>
<td>54.341</td>
<td>49.204</td>
<td>55.461</td>
<td>40.266</td>
<td>35.810</td>
<td>37.887</td>
</tr>
</tbody>
</table>

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


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