2023 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 17–20, 2023, ROME, ITALY

LIGHTWEIGHT IMAGE INPAINTING BY STRIPE WINDOW TRANSFORMER WITH JOINT ATTENTION TO CNN

Anonymous

ABSTRACT

Image inpainting is an important task in computer vision. As admirable methods are presented, the inpainted image is getting closer to reality. However, the result is still not good enough in the reconstructed texture and structure based on human vision. Although recent advances in computer hardware have enabled the development of larger and more complex models, there is still a need for lightweight models that can be used by individuals and small-sized institutions. Therefore, we propose a lightweight model that combines a specialized transformer with a traditional convolutional neural network (CNN). Furthermore, we have noticed most researchers only consider three primary colors (RGB) in inpainted images, but we think this is not enough. So we propose a new loss function to intensify color details. Extensive experiments on commonly seen datasets (Places2 and CelebA) validate the efficacy of our proposed model compared with other state-of-the-art methods.

Index Terms— HSV color space, image inpainting, joint attention, stripe window, transformer

1. INTRODUCTION

Image inpainting has been studied by many researchers for several years. The main goal of image inpainting is to fill up the realistic pixels in the missing region of the image and this can be applied to object removal and photo restoration. To achieve realistic results, we need to consider the following two important points: 1) the continuity of adjacent textures; 2) visually reasonable structure. All the proposed methods target at the above two points to solve the problem, such as the traditional diffusion method, patch matching method and current methods (CNN and GAN). However, they still face some limitations because convolution-based CNN has a narrow receptive field and hence it cannot get global information for the whole image. Without global information of the whole image, it is hard to repair the key edge and lines within the scene. To address this, some researchers proposed methods that utilize auxiliary information for structure recovery, e.g., edge connect (EC) [1]. On the other hand, some researchers proposed an attention mechanism-based model using attention scores compared with each patch to obtain global information. Suvorov *et al.* [2] utilized the Fast Fourier Convolution (FFC) to encode features in the frequency domain with global receptive fields for resolution-robust inpainting. Although these methods have improved the overall repair results, they require a huge computational cost. Furthermore, in recent years, transformers have also been used in the inpainting field due to their wider receptive fields than CNNs and better inpainting at low resolutions. However, transformers require a significant amount of computer memory. Therefore, it inspired us to design a lightweight transformer block with stable repair effects.

Specifically, we referred to the CSWin transformer [3] which used stripe window self-attention to replace traditional full self-attention. The stripe window self-attention mechanism computes self-attention parallel to horizontal and vertical stripe cross-windows. Each stripe is obtained by dividing the input feature into constant-width stripes. In this way, we can achieve global attention with limited computational cost. Then we redesigned the transformer block to enhance its repair performance.

The consistency of color is another important factor to judge the quality of an image. It is easy to discern the difference between inpainted image and original image by the human eye even if there is only a small deviation in the color. While most existing methods only deal with the basic primary colors, we believe that this is not enough. If we can quickly improve color consistency in the early stage of training, the repair performance can be improved. Therefore, we transform the inpainted image to the HSV color space and compare it with the input image. In follow-up experiments, our method is confirmed to be effective.

The rest of the paper is organized as follows. In Section 2, we introduce the previous and state-of-the-art inpainting methods. Then we present our proposed method and loss function in Section 3. In Section 4, we exhibit our training details, experiment results, inpainting images, and ablation studies. At last, the conclusions are drawn in Section 5. The major contributions of this work are as follows:

- We propose a stripe window self-attention transformer with an efficient local enhancement position encoding. Then we redesign the transformer block to make the result better than the original method.
- We suggest joint attention from global layers to local

layers, connecting the two layers to enhance the overall consistency of repair results.

- We propose a new HSV loss focused on color consistency in the early stage.
- In the common dataset including Places2 and CelebA, we conduct extensive experiments to confirm that our proposed model is better than other advanced methods.

2. RELATED WORK

Traditional inpainting. Traditional inpainting methods can generally be divided into two categories: diffusion-based methods and patch-matching methods. Diffusion methods disseminate texture content from the known region to the missing region using one or multi-curve information. However, this method tends to blur inpainting results for large masks. Patch-matching [4] methods, on the other hand, use approximate nearest-neighbor to find the nearest-neighbor region of the specified region and then selected the most similar region to fill in and complete the image inpainting. However, this method can be computationally expensive.

Deep learning based inpainting. With the increasing availability of advanced hardware technology, CNN-based deep learning models have emerged as the predominant approach for image inpainting. Several deep models have been proposed in this field, including Shift-Net [5] proposed by Yan *et al.* and more recent models that leverage additional information such as edge information. For example, Nazeri *et al.* proposed Edgeconnect [1] and Yu *et al.* proposed DeepFill-V2 [6] which used Canny edge detection to generate edge images. Zeng *et al.* [7] proposed CRFill, which utilized auxiliary contextual reconstruction loss to encourage the generator network to borrow appropriate known regions as references for filling in a missing region. While these methods have been shown to be effective in inpainting images with complex structures such as buildings and interior spaces, they require additional stages or parameters during training. In our proposed method, we also utilize edge information, but we avoid the need for additional parameters in the model.

On the other hand, some researchers have utilized self-attention mechanisms to improve texture inpainting, such as CA proposed by Yu *et al.* [8] and HiFill proposed by Yi *et al.* [9]. These methods compute complex attention scores to identify the most similar texture to be used in filling the missing region, and generally outperform other methods in terms of texture quality. In our proposed method, we have redesigned the attention module and incorporated wide attention to the local receptive field to enable attention sharing.

Vision transformer. In recent years, the use of transformer models in computer vision has gained popularity. He *et al.* proposed the Vision Transformer (ViT) [10], which made the transformer architecture applicable to computer vision tasks. Since then, more novel transformers have been introduced, such as Dong *et al.*'s CSWin transformer [3], and some have been applied in image inpainting, such as Zheng *et al.*'s TFill [11]. Transformers are able to inpaint plausible textures for large missing regions by using their special attention mechanism. However, they require more computing resources than traditional convolutional neural networks due to their wider receptive field. In our proposed method, we have redesigned the basic transformer architecture and utilized a stripe window to divide the feature map, reducing the amount of computations and achieving better repair effects.

To summarize, this paper proposed a novel stripe window-based transformer framework for image inpainting, and enhanced it with joint attention local CNN layers. Our model focuses on the global Stripe Window Multi-Head (SWMH) transformer and CNN-based local layer. We process the global and local layer in parallel and then share the same attention information between them. In the end, we use four simple up-samples to obtain the final inpainting result.

3. METHODOLOGY

Overview. The whole model of our proposed approach is shown in Fig. 1. Given a masked image $\mathbf{I_m}$ and a binary mask $\mathbf{M}$ which are both in 256×256, we concatenate them, and pass them through three downsampling CNN layers. After we downsample input image, we split the channel to global layer (i.e., SWMH transformer) and local residual in residual dense block (RRDB) [13] layer, where we use joint attention between the global and local layers. Each RDB block in RRDB has four consecutive Conv-ReLU. At last, we concatenate the features from both channels and then go through three upsample layers to get the inpainted image $\mathbf{I_{out}}$.

3.1. Stripe Window Multi-Head (SWMH) Transformer

The overall global layer of SWMH transformer is shown in Fig. 1. The input of the global layer is a feature map with size of $\mathbf{H}\times\mathbf{W}\times\mathbf{C}$, where $\mathbf{H}$ and $\mathbf{W}$ are 32 after downsampling and the channel is 128 after the split. There are four SWMH transformer blocks in our global layer. Each block has its own multi-head and stripe window (sw) to reduce the amount of calculation. We set multi-head to $2, 4, 8, 16$ and sw to $4, 8, 16, 32$ for the four blocks by default. The first three blocks are SWMH transformer blocks that split their channel into horizontal and vertical stripes, and then split their channel with their own multi-head again. The sw will split $\mathbf{H}$ or $\mathbf{W}$ depending on the choice of horizontal stripes or vertical stripes. In contrast to general multi-head self-attention (MHSA), our stripe window multi-head self-attention (SWMH-SA) combines multi-head and sw to greatly reduce computational complexity and achieve better inpainting effects. The last block of the SWMH transformer uses full attention because the sw in the fourth block is 32, which means the stripe window covers the whole image.

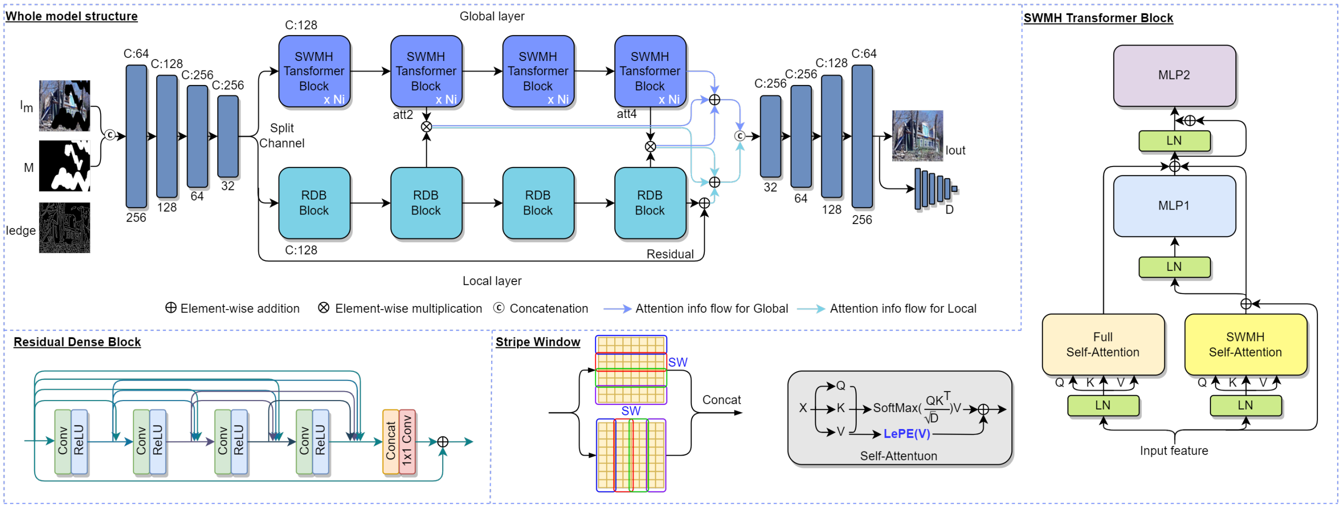

Fig. 1: The overview of our proposed model. The **whole model structure** shows the framework of our proposed model and the details of the joint attention between Global layer and Local layer. The input images only include I_m and M. The I_{edge} will be used in the loss function and generated by Canny [12] before training. Moreover, the right side shows the **SWMH Transformer Block**. D is the normalization factor before softmax, which makes the similarity between pixels become more stable. At last, the **Residual Dense Block** in the local layer is shown at the bottom left corner of the whole model.

SWMH Transformer Block. The structure of SWMH Transformer Block is also shown in Fig. 1. We redesign the self-attention wiring, moving it from the first feed-forward to the beginning because we hope our self-attention block will not be influenced by the SWMH-SA. SWMH Self-Attention and Full Self-Attention will be trained from different receptive fields and then connected together with the residual link. We also add locally-enhanced positional encoding (LePE) in the transformer block to augment the positional encoding and refer to [3] to add the LePE at the end of the transformer block but not the middle, shown on the right side of Fig. 1. We found that self-attention needs to be calculated multiple times to get better attention information. We set the $\mathbf{N_i}$ to denote the number of repetitions.

3.2. Joint attention

We concatenate global and local layers to jointly focus on the information with different receptive fields. We expect our inpainting results to be the admixture of different receptive fields, not only single receptive field. So we collect attention from the second and fourth SWMH transformer blocks and multiply it by the corresponding RDB blocks. At last the two mixed receptive fields are added to the respective last block of the global and local layers to achieve joint attention.

3.3. Loss Function

Most loss functions we adopt in this paper are the same as [1, 14, 15]. And we also use other losses including perceptual loss, Edge loss and HS loss which we proposed in this work. First, the $\mathbf{I_{out}}, \mathbf{I_{GT}}$ indicate predicted images and the ground truth, respectively. We enhance the structure of the inpainting image by using Edge loss which is $L_{edge} = \frac{1}{n}\sum_{i=1}^{n}||(I_{out} \odot M_{edge} - I_{GT} \odot M_{edge})||_2^2$, where n represents the number of pixels in the image, and $\mathbf{M_{edge}} = (1 - I_{edge}) + 10 * I_{edge}$, which can be seen as an edge mask to accentuate the edge structure. The $\mathbf{I_{edge}}$ is the image obtained from Canny edge detection [12].

In order to improve the quality of the inpainting model, we use Perceptual loss to measure the similarity between images. We also use the mask on feature map to let Perceptual loss only focus on visible regions. The VGG-19 based perceptual loss would force the model to generate images semantically closer to the ground truth, but we notice our inpainting results have checkerboard artifacts. According to [15], checkerboard artifacts are usually caused by deconvolution and using Style loss can remove this artifact. Therefore, we use the same Style loss as [15] in our total loss.

Besides focusing on texture and structure, we believe that color is as important as both. So we proposed the HS loss to measure the similarity between colors, which can be formulated as follows:

$$\begin{aligned}
&L_{HS} = \lambda_{HS} * \frac{1}{n}\sum_{i=1}^{n}||(HS_{out} - HS_{GT})||_2^2, \\
&L_{HS\_edge} = \frac{1}{n}\sum_{i=1}^{n}||(HS_{out} \odot M_{edge} - HS_{GT} \odot M_{edge})||_2^2, \\
&L_{HS\_T} = \lambda_{HS} * L_{HS} + \lambda_{HS\_edge} * L_{HS\_edge},
\end{aligned} \tag{1}$$

where $\lambda_{HS} = 10$ and $\lambda_{HS\_edge} = 100$ by default. Here, **HS** means $Hue, Saturation$ in HSV color space but we do not use $Value$ in the HS loss because brightness (intensity) can easily be included by other losses. If we still use the $Value$ in HS loss it will even affect our inpainting results. The L_{HS_edge} uses the edge mask, and we set it to have larger weight to enhance the boundary. We will demonstrate this in ablation experiments.

The adversarial loss includes the discriminator loss L_D and the generator loss L_G. The adversarial loss can be indicated as

$$\begin{aligned} &L_D = -\mathbb{E}_{I_{GT}}[logD(I_{GT})] - \mathbb{E}_{I_{outM}}[logD(I_{out}) \odot (1-M)] \\ &- \mathbb{E}_{I_{outM}}[log(1 - D(I_{out})) \odot M], \\ &L_G = -\mathbb{E}_{I_{out}}[logD(I_{out})], L_{adv} = L_D + L_G + \lambda_{GP} L_{GP}, \end{aligned} \tag{2}$$

where the PatchGAN [16] based discriminator is written as **D** and our proposed model can be seen as the generator **G**. The $L_{GP} = \mathbb{E}_{I_{GT}} || \bigtriangledown_{I_{GT}} D(I_{GT})||^2$ is the gradient penalty and $\lambda_{GP} = 1e-3$. We include all losses above as the total loss $\mathbf{L_{total}}$:

$$\begin{aligned} L_{total} = \lambda_{L1} L_1 + \lambda_{edge} L_{edge} + \lambda_{perc} L_{perc} \\ + \lambda_{style} L_{style} + L_{HS\_T} + \lambda_{adv} L_{adv}, \end{aligned} \tag{3}$$

where $\lambda_{L1} = 10$, $\lambda_{edge} = 10$, $\lambda_{perc} = 0.1$, $\lambda_{style} = 250$, and $\lambda_{adv} = 10$. The above loss weights are empirically set by experiments.

4. EXPERIMENTS

4.1. Datasets

To show the inpainting effectiveness of our proposed model, we conduct experiments on Places2 dataset. For Places2, we randomly chose 20k images from the original dataset as the training set, 5k images as the validation, and used about 4k images as the test. We use less data and the lightweight model to show our proposed approach has better robustness than other state-of-the-art huge-parameter models. For CelebA dataset, we split the dataset into 8:1:1 for training, validation and test. For all of the images in above two datasets, we only train and test them with image size 256×256. For other comparison methods, we use their provided pretrained models to perform the test on the same dataset as we did.

4.2. Reference State-of-the-Art

We compare the proposed model with other state-of-the-art (SOTA) methods, which include PatchMatch (PM) [4], Contextual Attention (CA) [8], Shift-net (SN) [5], Partial Convolutions (PC) [15], Gated Convolution (DeepFill-v2) [6], Contextual Residual Aggregation (HiFill) [9], Imputed Convolution (Iconv) [17], Aggregated contextual transformations (AOT-GAN) [18], Auxiliary Contextual Reconstruction (CR-Fill) [7], Bridging Global Context Interactions (TFill) [11].

4.3. Quantitative Comparisons

In Table 1, we utilize PSNR, SSIM [19] and LPIPS to assess the performance of all compared methods and our proposed approach on the two datasets with irregular masks of different masking rates. The model parameters are also shown beside each method, where the results are tested by ourselves. For two datasets, our proposed method can defeat most of compared methods in terms of these three evaluation metrics. Among them, LPIPS [20] is considered a better metric than other metrics in the inpainting field, because LPIPS used perceptual distance to compare high-level information which will be better than other low-level metrics. Hence, we also use this metric to compare the performance for all methods, which shows the robustness of our proposed model. On the other hand, our training images and steps are also less than most methods, so we can see the proposed method is effective from Table 1.

4.4. Qualitative Comparisons

We show the qualitative inpainting results of Places2 and CelebA in Fig. 2. Compared with other methods, our proposed model can reconstruct similar or even more clear textures. We notice our inpainting results are slightly blurred when we focus more on the transformer and less on CNN. In the future, we will set restrictions on the local layers so that local information will not be ignored. Furthermore, our architecture is a lightweight model, which means we do not need lots of parameters, and still can achieve similar results compared to those larger models. Note that both our training data and steps are less than other methods.

4.5. Ablation Study

To confirm our proposed module and new loss function are useful in the proposed architecture, we separately test them in the ablation experiments. We test the stability of the SWMH transformer and the redesign in Table 2. We retrained the CSWin transformer without redesign and original transformer [10] separately and compared them with our redesigned SWMH transformer. For the results shown in Table 2, our proposed approach has the best PSNR, SSIM, and LPIPS.

We also conduct experiments for HS loss in Table 2. We noticed the Value (V) of HSV can easily be learned in L_1 and other losses. If we still consider V in L_{HS}, it will influence the balance of the inpainting result, as shown in the table. We show the color deviation between with and without L_{HS} at early training steps in Fig. 3. We can see the color of the inpainting results in the early 50 training steps, which shows the one with L_{HS} is more close to the ground truth than without L_{HS}, and the known region and the missing region are more consistent when using L_{HS}.

At last, in order to confirm the joint attention with local layer is effective, we remove the whole local layer and only keep the global layer (i.e., w/o RDB). We can see the inpainting results become worse without local layers.

Table 1: Quantitative evaluation of inpainting on Places2 and CelebA datasets. We report *Peak signal-to-noise ratio* (PSNR), *structural similarity* (SSIM) and *Learned Perceptual Image Patch Similarity* (LPIPS) metrics. The ▲ denotes larger, and ▼ denotes lesser of the parameters compared to our proposed model. (**Bold** means the 1st best; Underline means the 2nd best)

		Places2 | CelebA												
		PSNR ↑						**SSIM ↑**						**LPIPS↓**
	Parameters x10^6	mask												
		5%~10%	10%~20%	20%~30%	30%~40%	40%~50%	50%~60%	5%~10%	10%~20%	20%~30%	30%~40%	40%~50%	50%~60%	5%~60%
PM [2009]	-	22.873 | 21.440	21.523 | 21.464	19.780 | 20.582	17.204 | 18.392	17.397 | 17.531	14.921 | 14.165	0.937 | 0.928	0.894 | 0.909	0.882 | 0.869	0.750 | 0.817	0.728 | 0.773	0.594 | 0.661	- | -
CA [2018]	3 ▼	30.698 | 34.559	26.575 | 29.554	26.323 | **29.214**	22.637 | 25.107	21.899 | 24.317	20.366 | 22.454	0.962 | 0.955	0.910 | 0.928	0.903 | 0.921	0.816 | 0.822	0.775 | 0.811	0.710 | 0.760	0.1831 | 0.1226
SN [2018]	55 ▲	24.431 | 20.753	23.057 | 19.320	22.957 | 18.757	22.685 | 17.176	20.598 | 15.718	18.306 | 15.475	0.893 | 0.822	0.868 | 0.818	0.842 | 0.762	0.807 | 0.673	0.708 | 0.579	0.587 | 0.537	0.2221 | 0.2647
PC [2018]	49 ▲	25.566 | 24.902	23.429 | 23.218	23.475 | 23.392	24.226 | 22.359	23.275 | 21.005	22.661 | 22.494	0.879 | 0.859	0.845 | 0.846	0.834 | 0.844	0.829 | 0.811	0.803 | 0.765	0.768 | **0.793**	0.2182 | 0.1924
DeepFill v2 [2019]	4 ▼	32.741 | 33.282	28.329 | 28.667	27.015 | 28.634	24.117 | 25.128	23.391 | 24.515	21.713 | 22.563	0.966 | 0.972	0.921 | 0.924	0.904 | 0.829	0.835 | 0.865	0.799 | 0.815	0.732 | 0.776	0.1284 | 0.1885
HiFill [2020]	3 ▼	27.128 | -	22.391 | -	21.906 | -	18.282 | -	17.241 | -	15.704 | -	0.930 | -	0.825 | -	0.804 | -	0.671 | -	0.580 | -	0.488 | -	0.2506 | -
Iconv [2020]	30 ▲	27.671 | 27.174	23.629 | 27.174	23.179 | 26.729	20.382 | 23.712	19.396 | 22.841	18.313 | 21.476	0.933 | 0.877	0.839 | 0.877	0.822 | 0.863	0.707 | 0.782	0.628 | 0.719	0.552 | 0.666	0.3810 | 0.2517
AOT-GAN [2020]	15 ▲	31.078 | 30.970	28.231 | 28.558	**27.947** | 28.389	24.600 | 25.181	23.741 | 24.539	22.184 | 22.827	0.950 | 0.946	0.913 | 0.915	0.907 | 0.909	0.832 | 0.854	0.791 | 0.821	0.728 | 0.772	0.1482 | 0.1104
CRFill [2021]	4▼	32.679 | 32.526	27.806 | 27.443	27.339 | 27.099	23.805 | 23.095	22.938 | 22.308	21.418 | 20.522	0.964 | 0.966	0.914 | 0.916	0.906 | 0.908	0.828 | 0.832	0.787 | 0.791	0.729 | 0.733	0.1925 | 0.1379
TFill [2022]	15 ▲	**33.191** | **35.143**	28.717 | **29.269**	27.42 | 28.664	24.43 | 25.651	23.684 | 24.517	21.915 | 22.86	**0.968** | **0.975**	0.922 | **0.929**	0.911 | 0.915	0.842 | 0.863	0.803 | 0.816	0.728 | 0.775	0.1331 | **0.0972**
Ours [2023]	6	31.175 | 31.782	**28.718** | 28.849	27.753 | 28.708	**24.842** | **25.907**	**24.127** | **24.616**	**22.866** | **22.916**	0.944 | 0.947	**0.923** | 0.924	**0.912** | **0.923**	**0.849** | **0.871**	**0.804** | **0.822**	**0.734** | 0.782	**0.1217** | 0.1032

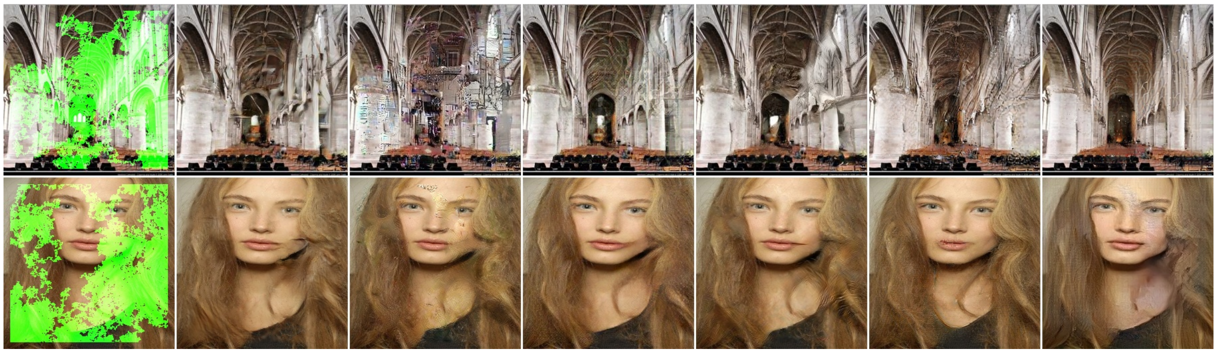

Fig. 2: Qualitative results of Places2 (upper half) and CelebA (lower half) datasets among SOTA methods. From left to right: Masked image, DeepFill-v2 [6], Iconv [17], AOT-GAN [18], CRFill [7], TFill [11], and Ours. Zoom-in for details.

Table 2: Ablation study of RDB, HS loss and SWMH transformer with size 256×256 images on Places2 dataset.

w/ RDB	w/ ViT	w/ CSWin	w/ SWMH	w/ HS_edge	w/ HSV	w/ HS	**PSNR↑**	**SSIM ↑**	**LPIPS↓**
x	x	x	✓	✓	x	✓	25.3351	0.7922	0.1400
✓	✓	x	x	✓	x	✓	25.7935	0.8072	0.1242
✓	x	✓	x	✓	x	✓	26.1027	0.8377	0.1221
✓	x	x	✓	x	x	x	26.2786	0.8459	0.1212
✓	x	x	✓	x	x	✓	26.3444	0.8422	0.1193
✓	x	x	✓	✓	✓	x	26.4757	0.8541	0.1184
✓	x	x	✓	✓	x	✓	**26.5801**	**0.8611**	**0.1156**

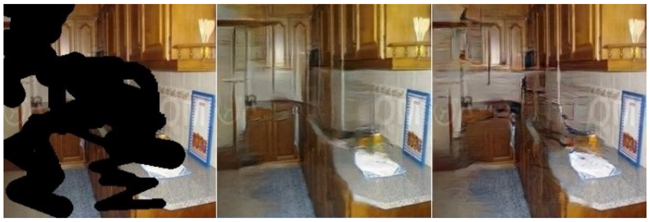

Fig. 3: Ablation study of color deviation on inpainted images. From left to right: Masked images, w/o L_{HS_T} loss, and w/ L_{HS_T} loss.

4.6. Object removal

Moreover, we demonstrate our model has practical applications in Fig 4. In this figure, we show object removal and background inpainting results.

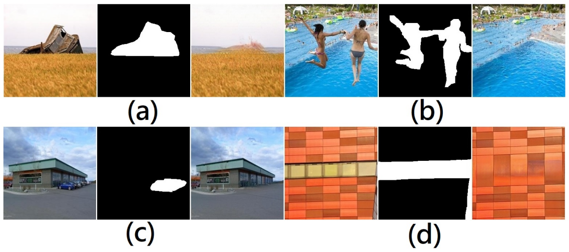

Fig. 4: Object removal (size 256×256) results. In (a)-(d), from left to right: Ground-truth image, mask, object removal result.

5. CONCLUSION

In this paper, we propose a lightweight joint attention transformer architecture. We use transformer-based architecture to get wide receptive field information and cooperate with local layers with RRDB by joint attention with each other. Our proposed HS loss can stabilize the colors in early training steps and eventually further improve the inpainting performance. We refer to the CSWin transformer and proposed the SWMH transformer block to not confuse the two self-attentions and achieve significant improvements. Our experiments demonstrate that the proposed model using small amount of parameters can still generate similar or even better inpainting results than other SOTA methods. Those large models do have an advantage in details but not every researcher has enough hardware support. Therefore, we propose this approach to demonstrate small models are also able to compete with large models.

6. REFERENCES

[1] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi, “Edgeconnect: Generative image inpainting with adversarial edge learning,” *arXiv preprint arXiv:1901.00212*, 2019.

[2] 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 *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2149–2159.

[3] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo, “Cswin transformer: A general vision transformer backbone with cross-shaped windows,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12124–12134.

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman, “Patchmatch: A randomized correspondence algorithm for structural image editing,” *ACM Trans. Graph.*, vol. 28, no. 3, pp. 24, 2009.

[5] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan, “Shift-net: Image inpainting via deep feature rearrangementperceptual,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 1–17.

[6] 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*, 2019, pp. 4471–4480.

[7] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M Patel, “Cr-fill: Generative image inpainting with auxiliary contextual reconstruction,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14164–14173.

[8] 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 conference on computer vision and pattern recognition*, 2018, pp. 5505–5514.

[9] 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*, 2020, pp. 7508–7517.

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick, “Masked autoencoders are scalable vision learners,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16000–16009.

[11] Chuanxia Zheng, Tat-Jen Cham, Jianfei Cai, and Dinh Phung, “Bridging global context interactions for high-fidelity image completion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11512–11522.

[12] Lijun Ding and Ardeshir Goshtasby, “On the canny edge detector,” *Pattern recognition*, vol. 34, no. 3, pp. 721–725, 2001.

[13] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018, pp. 0–0.

[14] 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*, 2022, pp. 11358–11368.

[15] 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)*, 2018, pp. 85–100.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[17] Håkon Hukkelås, Frank Lindseth, and Rudolf Mester, “Image inpainting with learnable feature imputation,” in *DAGM German Conference on Pattern Recognition*. Springer, 2020, pp. 388–403.

[18] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo, “Aggregated contextual transformations for high-resolution image inpainting,” *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[20] 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 conference on computer vision and pattern recognition*, 2018, pp. 586–595.