1 Introduction

This supplementary material presents the details of our network architectures in Section 2, more details about Laplace attention in Section 3 and more comparison results in Section 4.

2 Details of Network Architecture

Our proposed hourglass attention network (HAN) is built upon the framework of generative adversarial net. Specifically, the proposed HAN plays a role as a generator in the framework, and we adopt a PatchGAN[8] with the spectral normalization [5] network as our discriminator. We provide the details of the architectures of HAN and the PatchGAN in Table 2 and Table 1, respectively. Besides, "Resblock" and "Feed forward" of the hourglass attention structure details are shown in Figure 1.

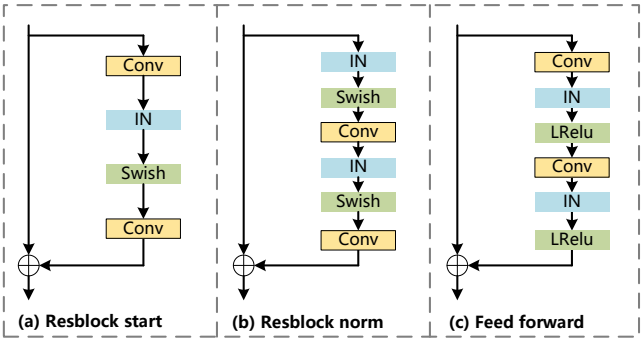

Fig. 1. Illustration of the Resblock and Feed forward block of attention block used in our model. (a) The Resblock used in the first layer. (b) The Resblock used in the other layers. (c) The Feed forward block of the proposed hourglass attention structure. "IN" denotes the instance normalization, "LRelu" is the LeakyReLU activation function and "Swish" represents the Sigmoid Linear Unit [3].

3 Details of Laplace Attention

The proposed Laplace attention can be seen as adding relative position encoding on top of vanilla multi-head self-attention. Specifically, after feeding a masked image into the CNN encoder, we obtain a feature map $F \in \mathbb{R}^{H \times W \times C}$. Then suppose the patch size is s, we can divide the feature F into $\frac{HW}{s^2}$ feature patches of size $s \times s \times C$. Therefore, in the calculation of attention, we will get a similarity matrix $M_s \in \mathbb{R}^{\frac{HW}{s^2} \times \frac{HW}{s^2}}$. Next we will can calculate the corresponding l_1 distance

Table 1. Details of the PatchGAN discriminator[8] with Spectral Normalization[5]. "SN-Conv2d" denotes a 2D convolution operator with spectral normalization to stabilize GAN training

Module Name	Filter Size	Channels	Stride	Nonlinearity
SN-Conv2d	4	64	2	LeakyReLU(0.2)
SN-Conv2d	4	128	2	LeakyReLU(0.2)
SN-Conv2d	4	256	2	LeakyReLU(0.2)
SN-Conv2d	4	512	1	LeakyReLU(0.2)
SN-Conv2d	4	1	1	Sigmoid

Table 2. Details of the proposed hourglass attention network (HAN). "Conv2d" denotes the 2D convolution layer. "ResBlock" represents a convolution block consists of 2 convolutions and a skip connection [2]. "HA" denotes the proposed hourglass attention structure. "UpSample" represents the bilinear interpolation. Final we show whether and what nonlinearity layer is used in the nonlinearity column

Module Name	Filter Size	Channels	Stride/UpFactor	Nonlinearity
Resblock	5	64	1	Swish
Resblock	3	128	2	Swish
Resblock	3	256	1	Swish
Resblock	3	256	1	Swish
Resblock	33	256	2	Swish
HA	33	256	1	LeakyReLU(0.2)
UpSample	null	256	2	null
Resblock	33	128	1	Swish
Resblock	33	128	1	Swish
UpSample	null	128	2	null
Resblock	33	64	1	Swish
Resblock	33	64	1	Swish
Conv2d	7	3	1	Tanh

matrix based on the matrix size $\frac{HW}{s^2} \times \frac{HW}{s^2}$. In addition, we additionally introduce a scalar $w \in \mathbb{R}^1$ to represent the effect of variance for each attention module.

For example, in our experiment, $H = 64$, $W = 64$ and $C = 256$. Assuming $s = 8$, the corresponding coordinates of each feature patch are shown in Figure 2. In order to perform the proposed Laplace attention, we need to compute the 2D Manhattan Distance from each patch of the $\frac{HW}{s^2} = 64$ patches to other patches (containing itself). Then we concatenate these results into a $\frac{HW}{s^2} \times \frac{HW}{s^2} = 64 \times 64$ matrix M_d. Next we multiply M_d with the learnable variance parameter $-|w|$ and add it ot the corresponding similarity matrix M_s. Finally, we pass the obtained results through softmax to obtain the corresponding attention scores $A \in \mathbb{R}^{64 \times 64}$, as:

$$A = \text{softmax}(M_s - |w| M_d) \tag{1}$$

In the experiment, the HAN model has four cases of s equal to 1, 2, 4, and 8, i.e., corresponding to four distance matrices of different sizes M_d. We can calculate these matrices in advance when constructing the model, and then just feed the corresponding matrices into the attention block in the process of the forward propagation respectively.

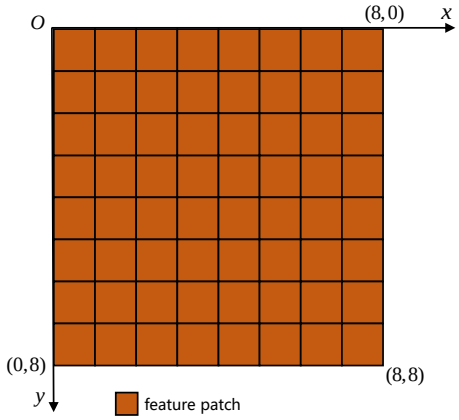

Fig. 2. Illustration of the axis defined in order to calculate the Manhattan Distance between feature patches.

4 More Results

We show more qualitative comparisons in Figure 3.

Fig. 3. Qualitative results with GC [7], RFR [4], DSN [6], DTS [1] and our models. The images in each of the three rows from top to the bottom are taken from CelebA-HQ, Paris street view, Places2 respectively. (Best viewed with zoom-in)

References

1. Guo, X., Yang, H., Huang, D.: Image inpainting via conditional texture and structure dual generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14134–14143 (October 2021)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
3. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
4. Li, J., Wang, N., Zhang, L., Du, B., Tao, D.: Recurrent feature reasoning for image inpainting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
5. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=B1QRgziT-
6. Wang, N., Zhang, Y., Zhang, L.: Dynamic selection network for image inpainting. IEEE Transactions on Image Processing **30**, 1784–1798 (2021). https://doi.org/10.1109/TIP.2020.3048629
7. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
8. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)