The Colorization Based on Self-Attention Mechanism and GAN

Jifeng Sun*
Institute of Information and Electronic South China University of Technology
Guangzhou, China
ecjfsun@scut.edu.cn

Yibin Lin
Institute of Information and Electronic South China University of Technology
Guangzhou, China

Shuai Zhao
Institute of Information and Electronic South China University of Technology
Guangzhou, China

2023 2nd International Conference on Artificial Intelligence and Computer Information Technology (AICIT) | 979-8-3503-0295-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/AICIT59054.2023.10277789

***Abstract*—Grayscale image colorization is a process of adding reasonable color information to an image, and converting grayscale images into color images is an important and difficult image processing task. The process of colorization is to predict the color information corresponding to the grayscale image by the colorization model. In this paper, the proposed multi-scale input adversarial generative network coloring model with multi-scale input is the colorization scheme based on self-attention mechanism and GAN is proposed in this paper. The experimental result on the colorization of grayscale cartoon images shows the effectiveness of the proposed scheme.**

Keywords-component; colorization; self-attention mechanism; transformer; GAN; upgrade network

I. INTRODUCTION

Grayscale image colorization is widely used in entertainment, medical science and military. The image colorization is an ill-posed problem, i.e. there are many colorization solutions to a special grayscale image. Different to an ordinary image, there are more colorization solutions, less texture and bigger block of the same color to a grayscale cartoon image than a image token from a nature scenery.

In the previous research [1,2,3,4]，Deshpande al. solved the problems of low definition, model corruption and long sampling time in the diversification of grayscale image colorization. On the other hand, Kumar al also made contribution to the solving of these problems. A grayscale image colorization based on multi-size input GAN is proposed as following:

In this paper, The colorization scheme based on self-attention mechanism and GAN is proposed in this paper, in order to the diversification and high quality of the colorization of grayscale cartoon images based on [4, 5], which consists of the low-definition colorization with Transformer and the upgrading high -definition colorization. This two-step scheme makes the diversity and high quality colorization of the grayscale image possible.

II. THE PRINCIPLE OF THE COLORIZATION BASED ON SELF-ATTENTION MECHANISM AND GAN

The self-attention mechanism originates from the vision of human and animal, in which there is non-uniformity in sensing the vision signal. In the machine learning based on the self-attention mechanism, the attention is in the importance parts of the input.

The human attention is divided as the autonomic attention and the non- autonomic attention, for example, A red apple and a book are put on the table, your attention may be attracted by the red apple, which is the autonomic attention. According to the non- autonomic attention, you may want to read book. A no weighted parameter pooling can be used to implement non- autonomic attention in the deep learning. Set q(query) is the query vector of the autonomic attention, v(value) is the input, k(key) is the key vector of the non-autonomic attention, the process of solving of k(key) from q(query) is shown Eq.(1) and Fig.1 .

$$\sum_{i=1}^{m} a(\mathbf{q},\mathbf{k}_i)\mathbf{v}_i \tag{1}$$

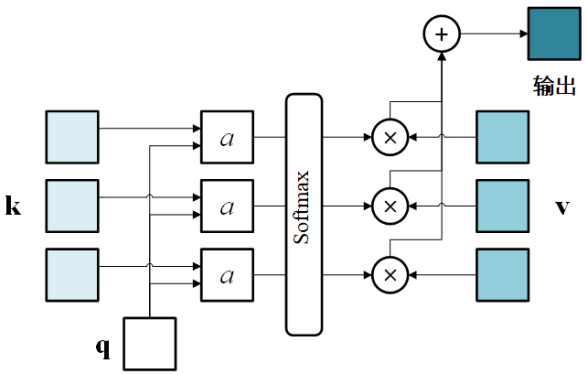

Figure 1. Attention mechanism

The attention model which consists of the self-attention and the non- self-attention is expressed by Eq.(1)

Where q represents the query vector corresponding to the self-attention, v the value vector corresponding to the input, k the key vector corresponding to the non-self-attention. q and k all possess the mean 0 and variance 1.

$$a(\mathbf{q},\mathbf{k}) = \mathrm{softmax}(\frac{\mathbf{q}^{\cdot}\mathbf{k}}{\sqrt{d}}),\quad \mathbf{q},\mathbf{k} \in \mathbb{R}^{d} \tag{2}$$

Eq.(2) can be written as the matrix form as Eq.(3).

$$\mathrm{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}(\frac{\mathbf{Q}\mathbf{K}^{\cdot}}{\sqrt{d}})\mathbf{V} \tag{3}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times d}, \mathbf{K} \in \mathbb{R}^{m\times d}, \mathbf{V} \in \mathbb{R}^{m\times v}$, A is an $n \times v$ matrix.

Qi in Q can expressed as the function parameter matrix Wi and input X as Eq(4)

$$\begin{aligned} \mathbf{Q}_i &= \mathbf{X}\mathbf{W}_i^q, \mathbf{K}_i = \mathbf{X}\mathbf{W}_i^k, \mathbf{V}_i = \mathbf{X}\mathbf{W}_i^v \\ \mathbf{H}_i &= \mathrm{A}(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) \\ \mathrm{MSA}(\mathbf{X}) &= \left[\mathbf{H}_1 \quad \cdots \quad \mathbf{H}_h\right]\mathbf{W}_o \end{aligned} \tag{4}$$

A. *The colorization model of the image*

For the image of the real world, the sky is blue, the grass is green. But the colors of the human cloths and the balloons are uncertain. The color diversity also occurs in the case of the cartoon. The scheme of the colorization model of the image shown in Fig.2 can be divided as two steps:

Step 1 Using the Transformer colorization network based on the self-attention mechanism to obtain the rough colorization result.

Step 2 Using upgrading network consists of encoder, medium layer and decoder to obtain the super-definition colorization result.

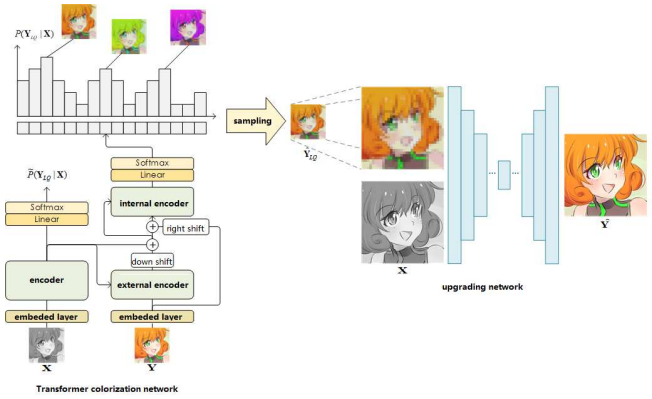

Figure 2. Diversity of the colorization model in the cartoon

B. *The preprocessing of the input*

In order to get the diversity of the colorization, some preprocessing will be done. Zhang[13] al. consider the colorization as the problem of classification and classified each pixel as 313 class. According to Zhang's method, we use the probability distribution to describe the color, which is easy to get the diversity result in the stochastic sampling about the probability.

C. *Transformer colorization network*

The colorization network based on Transformer and self-attention module are respectively shown in Fig.3 and Fig.4.

$$\begin{aligned} \left[\mathbf{q}_j, \mathbf{k}_j, \mathbf{v}_j\right] &= \mathrm{LN}(\mathbf{x}_i)\mathbf{W}_j^{qkv} \\ \mathbf{A}_j &= \mathrm{softmax}\left(\frac{\mathbf{q}_j \mathbf{k}_j^{\cdot}}{\sqrt{D_h}}\right) \\ \mathbf{h}_j &= \mathrm{SA}(\mathbf{x}_i) = \mathbf{A}_j \mathbf{v}_j \\ \mathrm{MSA}(\mathbf{x}_i) &= [\mathbf{h}_1, \ldots, \mathbf{h}_h]\mathbf{W}_o \\ \mathbf{x}_i' &= \mathrm{MSA}\left(\mathbf{x}_i\right) + \mathbf{x}_i \\ \hat{\mathbf{x}}_i &= \mathrm{MLP}\left(\mathrm{LN}\left(\mathbf{x}_i'\right)\right) + \mathbf{x}_i' \end{aligned} \tag{5}$$

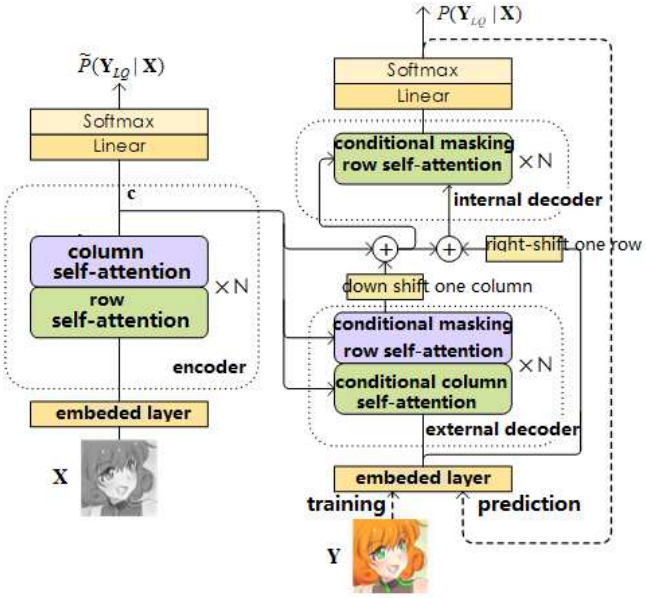

Figure 3. colorization network based on Transformer

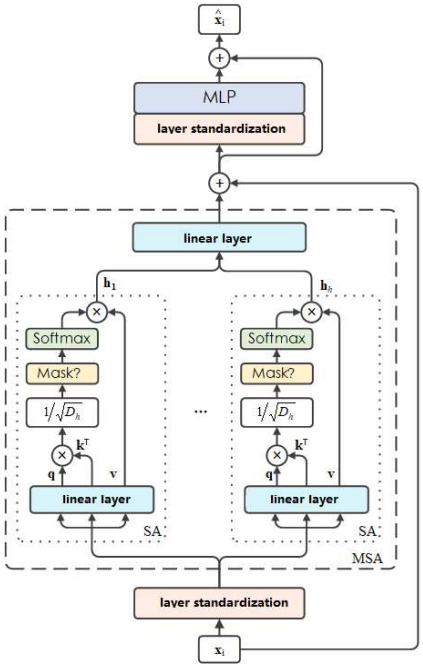

Figure 4. row or column self-attention module

D. upgrading network

The upgrading network and the discriminator network structure are respectively shown in Fig.5 and Fig.6.

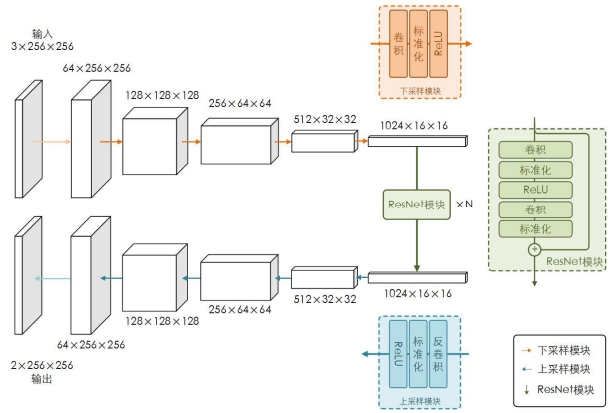

Figure 5. upgrading network structure

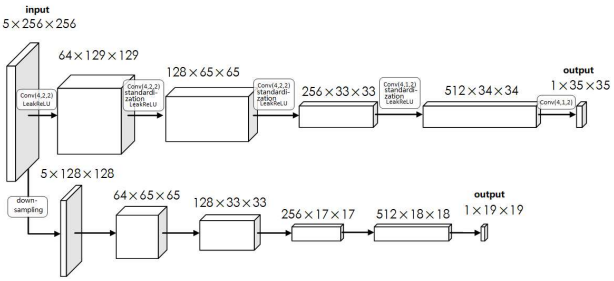

Figure 6. discriminator network structure

E. Loss function

The Loss functions are represented as Eq(8) and Eq(10) respectively.Ease of Use

$$L_c = -\left[\lambda_c \log P_c + (1-\lambda_c)\log \tilde{P}_c\right] \tag{7}$$

$$L_r = -\log P_r(\mathbf{Y} \mid \mathbf{X}_{LQ}) \tag{8}$$

$$\begin{aligned} G^* = \arg\max_D \min_G \mathrm{E}_{x,y}\left[\log D(x,y)\right] \\ +\mathrm{E}_x\left[\log\left(1-D\left(x,G(x)\right)\right)\right] \\ x \sim P\left(\mathbf{X}_{LQ}\right), y \sim P(\mathbf{Y}) \end{aligned} \tag{9}$$

$$L_r = L_{GAN} + \lambda_r L_{MSE} = L_{GAN} + \lambda_r \mathrm{E}_{x,y}\left[\left\|y - G(x)\right\|_2\right] \tag{10}$$

III. EXPERIMENT

Datasets are Anime Face Datasets[20], the subset of DANBOORU2018 and human face cartoon subset of DANBOORU2018[21],140000 images in all, whose sizes are 512×512, some of them given in Fig.9.

Workstation setting is RTX 3080 GPU\ Ubuntu 20.04\Tensorflow or pytorch. The parameters of Transformer are batch=4 and λ_c =0.99, For Optimized RMSProp algorithm, the regular learning rate is 3×10-4,, the times of iteration are 1.6×106. For the ascending network, batch is 1, λ_r in Eq.(10) is set as 10， For ADAM algorithm of $\beta_1 = 0, \beta_2 = 0.9$， the regular learning rate is 2×10-4，the times of iteration are 2.6×106.

13000 512×512 images with 8-bit color of dataset are down-samplied as 64×64 images with 3-bit color, and are used training dataset, 1000 of which are used as the test dataset Colorization experiment

Fig 7 shown the curve of the loss function of the training process for the Transformer colorization network.

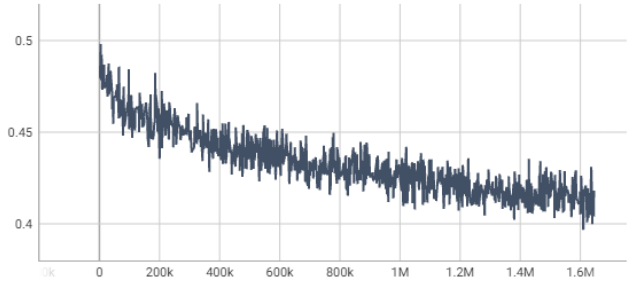

Figure 7. Curve of the loss function

Original images | Gray level images | Results 1 of diversity | Results 2 of diversity | Results 3 of diversity | Results 4 of diversity | Results 5 of diversity

Figure 8. Diversity experiment with Transformer colorization network

A. Experiment of the upgrading network

The training of the upgrading network deal with original image X, color-down-sampling image X_{LQ} Table 1 and Fig.9 show SSIM, PSNR and colorizing image of the ascending network respectively.

TABLE I. the upgrading result of the low quality images

	SSIM	PSNR
Before upgrading	0.6776	19.3983
Interpolation and up-sampling	0.9497	31.0157
After upgrading	0.9617	36.5282

TABLE II. the upgrading g result of the coarse colorization

	SSIM	PSNR
Before ascending	0.6058	16.4790

Interpolation and up-sampling	0.8635	25.7254
Before ascending	0.8800	20.3288

Figure 9. Colorization results before ascending and after ascending

B. *Colorization experiment*

The comparison experiment with VAE-MDN[1], cGAN[22], cINN[23] is done in order to identify the effectiveness of the proposed colorization scheme, the SSIM, PSNR and colorizing image result of which are respectively shown in Table 3 and Fig.13.

TABLE III. FID score of various colorization methods

method	cGAN[22]	VAE-MDN[1]	cINN[23]	Our method
result1	33.9337	25.8328	20.409	19.4743
result 2	33.9337	25.6193	20.4555	19.4016
result 3	33.9487	24.9107	20.4493	19.3829
result 4	33.9564	24.0664	20.2886	19.3639
result 5	34.0491	23.8004	20.4384	19.5790
result 6	34.0335	23.9653	20.4603	19.4926
mean	33.9758	24.6992	20.4169	**19.4490**

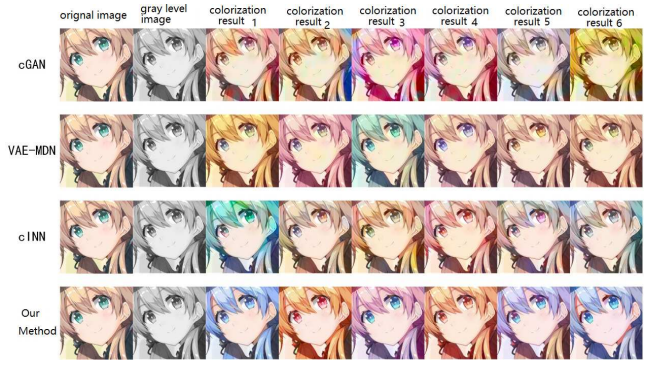

Figure 10. Results of various colorization methods

IV. CONCLUSION

The colorization scheme based on self-attention mechanism and GAN is proposed in this paper, which can realize the diversity of the colorization of the grayscale cartoon. The scheme can be divided as two steps.

The experimental result on the colorization of grayscale cartoon images shows the effectiveness of the proposed scheme

ACKNOWLEDGEMENT

The project is supported by NSFC No. 62071183. We thank the support of the National Natural Science Foundation of China. We thank Professor Lianwen Jin from South China University of Technology for his providing high performance workstations.

REFERENCES

[1] Deshpande A., Lu J, Yeh M. C, et al. Learning diverse image colorization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6837-6845.

[2] Liu M. Y., Breuel T., Kautz J. Unsupervised image-to-image translation networks[J]. Advances in neural information processing systems, 2017, 30.

[3] Guadarrama S., Dahl R., Bieber D., et al. Pixcolor: Pixel recursive colorization[J]. arXiv preprint arXiv:1705.07208, 2017.

[4] Vitoria P., Raad L., Ballester C.. Chromagan: Adversarial picture colorization with semantic class distribution[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2020: 2445-2454.

[5] Kumar M., Weissenborn D., Kalchbrenner N. Colorization transformer[J]. arXiv preprint arXiv:2102.04432, 2021.

[6] Lin Z, Feng M, Santos C N, et al. A structured self-attentive sentence embedding[J]. arXiv preprint arXiv:1703.03130, 2017.

[7] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[8] Paulus R, Xiong C, Socher R. A deep reinforced model for abstractive summarization[J]. arXiv preprint arXiv:1705.04304, 2017.

[9] Wolf T, Debut L, Sanh V, et al. Transformers: State-of-the-art natural language processing[C]//Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. 2020: 38-45.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[11] Wang T. C., Liu M. Y., Zhu J. Y., et al. High-resolution image synthesis and semantic manipulation with conditional gans[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8798-8807.

[12] Wang Z, Cun X, Bao J, et al. Uformer: A general u-shaped transformer for image restoration[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 17683-17693.

[13] Zhang R, Isola P, Efros A A. Colorful image colorization[C]//Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14. Springer International Publishing, 2016: 649-666.

[14] Ho J, Kalchbrenner N, Weissenborn D, et al. Axial attention in multidimensional transformers[J]. arXiv preprint arXiv:1912.12180, 2019.

[15] Wang H, Zhu Y, Green B, et al. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV. Cham: Springer International Publishing, 2020: 108-126.

[16] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[17] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[J]. arXiv preprint arXiv:1511.06434, 2015.

[18] Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization[J]. arXiv preprint arXiv:1607.08022, 2016.

[19] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[20] http://www.seeprettyface.com/mydataset_page3.html#anime

[21] https://gwern.net/danbooru2021#danbooru2018

[22] Cao Y, Zhou Z, Zhang W, et al. Unsupervised diverse colorization via generative adversarial networks[C]//Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18–22, 2017, Proceedings,

[23] Part I 10. Springer International Publishing, 2017: 151-166.

[24] Ardizzone L, Lüth C, Kruse J, et al. Guided image generation with conditional invertible neural networks[J]. arXiv preprint arXiv:1907.02392,