IMPROVING REFERENCE-BASED IMAGE COLORIZATION FOR LINE ARTS VIA FEATURE AGGREGATION AND CONTRASTIVE LEARNING

Shukai Wu[1] *Qingqin Wang*[1] *Shuchang Xu*[1,2,3,*] *Sanyuan Zhang*[1]

[1] College of Computer Science and Technology, Zhejiang University
[2] College of Information Science and Technology, Hangzhou Normal University
[3] Institute of Big data and Information Technology, Wenzhou University

ABSTRACT

The tremendous semantic discrepancy between the line art drawings without texture and the reference pictures containing rich color challenges current image-to-image translation models. Previous works attempt to establish cross-domain correspondence. However, they fail to capture more detailed features. A Reference-based Line art Translation Network (RLTN) is introduced with a Multi-level Feature Aggregation Module (MFAM) to improve the performance. The MFAM concentrates on more meaningful information for feature matching by utilizing the Multi-stream High Frequency Block (MHFB) and the Pixel-wise Correlation Block (PCB). We also employ the Channel-level Attention Block (CAB) and the Spatial-level Attention Block (SAB) for a better fusion of features. Moreover, a Style-based Contrastive Loss (SCL) is proposed to maintain the style similarity between the synthesized images and the reference examples. Experiments conducted on three datasets demonstrate the effectiveness of our model in producing more pleasing visual effects compared with state-of-the-art approaches.

Index Terms— Image-to-image Translation, Reference, Line Arts, Feature Aggregation, Contrastive Learning

1. INTRODUCTION

Benefiting from the development of deep learning, image-to-image translation techniques have been successfully applied into the fields including style transfer, image colorization, etc. Especially automatic image generation based on line arts has attracted continuous attention from the public recently, because it could provide users who have no painting experience with the freedom to create. However, when reference cases need to be considered simultaneously, the cross-domain difference challenges current models.

Despite the great success that previous works [1, 2] have achieved, these models fail to synthesize realistic photos while learning style from reference pictures. Approaches like [3, 4] disentangle images into the feature spaces of content and style (or attribute), then achieve image translation through feature recombination. Nevertheless, the local style information tends to be washed away in the encoding stage, and only the global style can be preserved. Reference-based methods [5, 6] attempt to establish dense correspondence between two different visual domains. They utilize attention techniques for better feature matching, bringing us amazing results. However, the features from different levels of the encoders are not fully employed, which means various contextual information may be ignored. How to improve the performance of the mapping process remains to be explored.

Fig. 1. The generated results using different references.

To alleviate the problems mentioned above, we propose an image-to-image translation model for reference-based image generation on line art drawings according to the previous work [6], namely **Reference-based Line art Translation Network (RLTN)**. With the purpose of making the most use of latent information from different stages of the encoders, a novel Multi-level Feature Aggregation Module (MFAM) is introduced to enhance feature representations. It can be observed that color is nonexistent in the black-and-white sketches. Thus, we believe an efficient way of building a robust cross-domain correlation is to focus on more significant features such as high-frequency components that the two domains all possess. Specifically, we inject a Multi-stream High Frequency Block (MHFB) and a Pixel-wise Correlation Block (PCB) into the MFAM for extracting the structural features related to edges and positions. Then, the Channel-level Attention Block (CAB) and the Spatial-level Attention Block

*Corresponding Author. E-mail address: w_sk@zju.edu.cn. This work is supported by the National Natural Science Foundation of China (LQ20F050011), the Plan Project of Wenzhou Municipal Science and Technology (R2020025), and Beijing Learnings Co., Ltd.

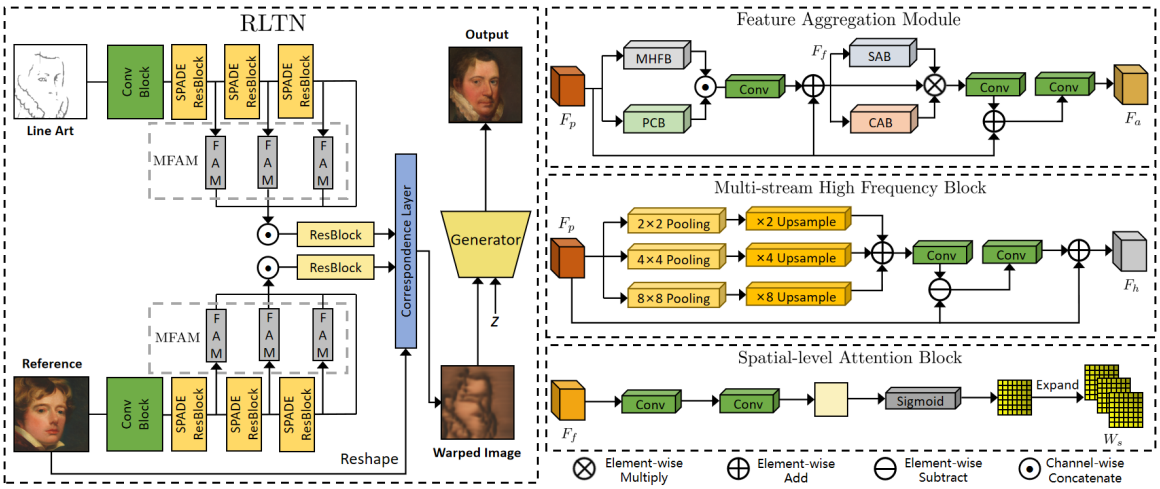

Fig. 2. The structure of RLTN. Our model takes a line art drawing and a reference picture as input, then generates a warped image by matching the high-dimensional aggregated features. The generator is responsible for producing the final output.

(SAB) make it possible to weigh the importance of feature maps. We fuse the deep features and leverage them for learning better correspondence. Furthermore, to strengthen style expression in the final generated pictures, a Style-based Contrastive Loss (SCL) is employed to guide the model for transferring more style. Experimental results verify the effectiveness of our model in generating vivid photos when considering references simultaneously, as shown in Fig. 1.

The main contributions can be summarized as follows:

- We introduce a Reference-based Line art Translation Network (RLTN) with a Multi-level Feature Aggregation Module (MFAM) to capture and fuse features from different stages utilizing attention techniques, which is conducive for finding cross-domain correspondence.
- A Style-based Contrastive Loss is presented to maintain the style similarity between the generated pictures and the reference examples, enhancing the style expression in the generating process.
- Experimental results on three public datasets demonstrate the effectiveness of our model compared with state-of-the-art works.

2. PROPOSED METHOD

2.1. Overview

In this section, we first display the structure of our Reference-based Line art Translation Network (RLTN). Then, we describe the details of the Multi-level Feature Aggregation Module (MFAM) and the Style-based Contrastive Loss (SCL). Fig.2 shows an overview of the RLTN. First, we adopt a content encoder to obtain latent representations for the line domain and use another style encoder for the reference domain. Next, the Multi-level Feature Aggregation Module (MFAM) extracts and combines features from different levels of the encoders. Then, we calculate the correlation matrix with the correspondence layer in [7] to get a warped image. Finally, the generator tries to synthesize a picture fusing the content of the line art drawing and the style of the reference example.

2.2. Multi-level Feature Aggregation Module

The MFAM enhances feature representations by applying individual Feature Aggregation Module (FAM) at different levels in the content and style encoders. Each FAM is composed of four blocks. The Multi-stream High Frequency Block (MHFB) focuses on the high-frequency components such as contours, while the Pixel-wise Correlation Block (PCB) concentrates on the relation between pixels. After concatenating the information from the above two blocks, we update the feature maps via employing the weight matrix obtained by the Spatial-level Attention Block (SAB) and the Channel-level Attention Block (CAB). It is worth noting that our PCB and CAB are borrowed from the previous works [8, 9], because their effects have been proven in many scenes. We will discuss the details of FAM in the following sections.

2.2.1. Multi-stream High Frequency Block

According to the idea in the frequency domain, an image can be composed of high-frequency and low-frequency components. For line drawings, there are only high-frequency components like edges. Thus, it is more reasonable to find the relation between the high-frequency features. A novel MHFB is introduced to address this. As shown in Fig. 2, given the input $F_p \in R^{C\times H\times W}$, we extract features by utilizing multi-stream pooling layers with different receptive field sizes of

2×2, 4×4 and 8×8. Then, the nearest neighbor interpolation method is used to restore the size of the down-sampling features to the original scale. At this point, we have obtained the low-frequency representations $F_i (i = 1, 2, 3)$, which contain the general features. The high-frequency components are emphasized by the following operation between F_p and F_i:

$$F_h = Conv_{3\times\ 3}(F_p - Conv_{3\times\ 3}(\sum_{i=1}^{3} F_i)) + F_p, \quad (1)$$

where $Conv_{3\times3}(\cdot)$ represents a convolutional layer with the filter size of 3×3.

2.2.2. Spatial-level Attention Block

The feature concatenation for the outputs of MHFB and PCB promotes the latent feature expression. Nevertheless, it is also necessary to weigh the importance of the new feature maps. Therefore, we reduce the dimension of the input feature maps $F_f \in R^{C\times H\times W}$ by applying two successive 3×3 convolutional layers. Subsequently, a sigmoid function is applied at each spatial position to form a weight matrix for all the points. Finally, we expand the matrix to match the scale of F_f. The weight matrix $W_s \in R^{C\times H\times W}$ is computed as:

$$W_s = Broadcast(\sigma(Conv_{3\times\ 3\times1}(Conv_{3\times3\times c/r}(F_f)))), \quad (2)$$

where $Conv_{3\times3\times c/r}(\cdot)$ represents a convolutional layer with the filter size of 3×3, and the output channel dimension is c/r. r is the ratio for channel reduction. $\sigma(\cdot)$ denotes the sigmoid function. $Broadcast(\cdot)$ expands the size via duplicating values along the spatial axis.

2.2.3. Feature Aggregation

We also use the CAB to get the weight matrix $W_c \in R^{C\times H\times W}$ in the channel level. The significance of each position in F_f is assigned through element-wise multiplication. The residual mechanism is for boosting the original feature representations. We gain the aggregated features by:

$$F_a = Conv_{3\times\ 3}(Conv_{3\times\ 3}(F_f \otimes W_s \otimes W_c) + F_p), \quad (3)$$

where $\otimes$ denotes the Hadamard product. Fusing the aggregated features from various levels allows us to obtain richer information for constructing better correspondence.

2.3. Style-based Contrastive Loss

Inspired by [10], to strengthen the style expression in deep neural networks, we raise a Style-based Contrastive Loss (S-CL) to promise the style similarity. It is defined as:

$$\mathcal{L}_{SCL} = max(0, 1 - \frac{\sum_{i=1}^{N} ||g(\phi_i(x_n)) - g(\phi_i(x_w))||_2}{m + \sum_{i=1}^{N} ||g(\phi_i(x_p)) - g(\phi_i(x_w))||_2}), \quad (4)$$

Table 1. The data distribution in our experiments.

Dataset	Train	Test	Total
CelebA-HQ [12]	24,000	6,000	30,000
ASCP [13]	14,224	3,545	17,769
MetFaces [14]	1,069	267	1,336

where x_n, x_p, and x_w denote the ground truth of the line art, the reference example, and the warped image, respectively. m stands for the margin to measure the feature distance. $g(\cdot)$ is the gram matrix and $\phi(\cdot)$ represents the pre-trained VGG [11] model. N is the total number of the layers ($relu2\_1$, $relu3\_1$, $relu4\_1$ and $relu5\_1$) for computing the difference.

3. EXPERIMENTS

3.1. Datasets

CelebA-HQ dataset [12] contains 30,000 high-quality face pictures. We obtain the edges with the contour extractor [15] on segmentation masks to construct line drawings. **ASCP datasest** [13] (Anime Sketch Colorization Pair dataset) consists of 17,769 pairs of cartoon sketches and color pictures. We use the original sketches. **MetFaces dataset** [14] is composed of 1,336 human face portraits from works of art in the Metropolitan Museum. The HED [16] algorithm is used to obtain the edges as line arts. Table 1 lists the data distribution in our experiments. All images are resized to 256×256. We adopt cosine similarity to search for the most similar picture for each sample under a resolution of 8×8.

3.2. Evaluation Metrics

We use **Fréchet Inception Distance (FID)** [17] to compute the distance between the distribution of generated data and reference (real) data. The lower the value, the closer the distribution. **Learned Perceptual Image Patch Similarity (LPIPS)** [18] is selected to get the perceptual distance between each pair of generated images and references. A lower score indicates higher perceptual similarity. Fifteen people are invited to give **User Preference Score (UPS)** on 100 images produced by each method for a user study. The values ranging from 0 to 10 are based on quality and similarity.

3.3. Implementation Details

The structure of the generator and the object function follows the settings in [6], and the SPADE ResBlock in the encoder refers to [1]. The weight of our Style-based Contrastive Loss is set to 10. We build our model with PyTorch and train it using a GeForce RTX 3090 GPU with one image a batch. Adam optimizer is adopted with β_1= 0.9 and β_2 = 0.999 respectively. Following TTUR [17], the learning rate is set to 10^{-4} for the generator and 2×10^{-4} for the discriminator.

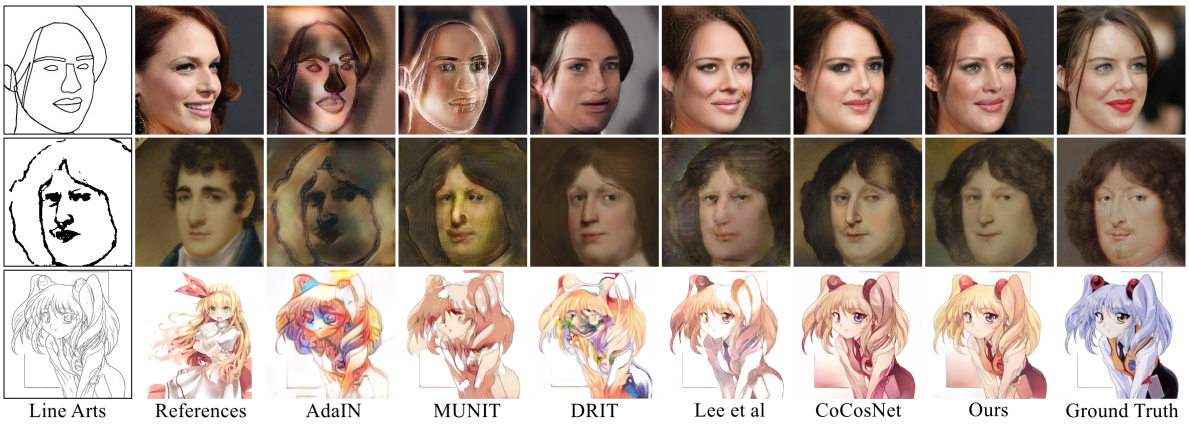

Fig. 3. The generated images produced by different methods.

Table 2. The quantitative results of different methods.

Method	CelebA-HQ		ASCP		MetFaces	
	FID	LPIPS	FID	LPIPS	FID	LPIPS
AdaIN [19]	186.24	0.561	114.84	0.458	249.93	0.606
MUNIT [3]	144.56	0.562	53.50	0.442	183.29	0.561
DRIT [4, 20]	57.43	0.466	44.58	0.445	118.28	0.552
Lee et al. [5]	54.11	0.387	34.01	0.435	165.18	0.574
CoCosNet [6]	24.86	0.332	27.02	0.433	76.77	**0.451**
Ours w/o $\mathcal{L}_{SCL}$	23.91	0.331	26.06	0.431	76.21	0.452
Ours w/o MFAM	24.37	0.332	26.77	0.434	76.63	0.453
Ours	**23.19**	**0.330**	**25.18**	**0.430**	**75.29**	**0.451**

Table 3. The results of the user study. Q and S denote quality and similarity, respectively.

Method	CelebA-HQ		ASCP		MetFaces	
	Q	S	Q	S	Q	S
AdaIN [19]	1.6	1.4	1.5	1.9	1.8	1.6
MUNIT [3]	3.1	2.5	3.2	2.8	3.1	3.4
DRIT [4, 20]	6.9	6.6	4.8	4.0	5.8	5.5
Lee et al. [5]	7.2	7.1	8.0	7.5	7.0	6.6
CoCosNet [6]	8.0	8.1	8.2	7.5	7.5	7.3
Ours	**8.1**	**8.3**	**8.4**	**7.7**	**7.7**	**7.4**

3.4. Compared with Existing Methods

To assess the performance of our model, we compare it with several classic and state-of-the-art approaches. Fig. 3 provides some generated cases. The results prove that our model surpasses other methods by producing more realistic photos with the most similar style to the reference examples. At the same time, the structure of the line arts is also well preserved. The style transfer method AdaIN [19] fails to transfer the line art drawings to high-quality pictures. The image-to-image translation models [3, 4, 20] produce photos that are not natural enough. The reference-based method [5] tends to bring unexpected shadows and disharmonic color distribution. CoCosNet [6] achieves better results compared with previous methods. However, the huge semantic discrepancy between two different domains limits the mapping ability, leading to failure in dealing with some details. To sum up, our model presents more stunning effects than other competitors.

Table 2 lists the quantitative results. Our method obtains the lowest FID, indicating that the distribution of our results is closer to that of the pictures in the reference domain. The best LPIPS also implies that we can keep a more similar style to the reference examples. The average scores of the user study are displayed in Table 3. We find that the participants are more inclined to choose our generated pictures. Our RLTN acquires the highest scores of quality and similarity on all the datasets, verifying the superiority of our model.

3.5. Ablation Study

We also conduct an ablation study to demonstrate the significance of the MHFB and the SCL. The roles of each part are explored individually. The values in Table 2 show that both of them are necessary. With the MHFB focusing on more valuable features and the SCL enforcing to pass more style, our model gains some improvements, which proves their essentiality in promoting the final performance.

4. CONCLUSION

In this paper, we introduce a Reference-based Line art Translation Network (RLTN) to produce images based on line art drawings with reference pictures. A Multi-level Feature Aggregation Module (MFAM) utilizing attention techniques is incorporated into our network for extracting better feature representations. We also construct a Style-based Contrastive loss (SCL) to promise the style similarity between the generated outputs and the reference examples. Experimental results on real-world datasets demonstrate the effectiveness of our model in synthesizing more realistic and vivid images.

5. REFERENCES

[1] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2337–2346.

[2] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[3] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[4] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International Journal of Computer Vision*, vol. 128, no. 10, pp. 2402–2417, 2020.

[5] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo, "Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5801–5810.

[6] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5143–5153.

[7] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen, "Deep exemplar-based video colorization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8052–8061.

[8] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.

[9] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[10] Paul Wohlhart and Vincent Lepetit, "Learning descriptors for object recognition and 3d pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3109–3118.

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

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[13] Taebum Kim, "Anime sketch colorization pair," `https://www.kaggle.com/ktaebum/anime-sketch-colorization-pair`, 2018, Accessed: 2021-07-21.

[14] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila, "Training generative adversarial networks with limited data," *arXiv preprint arXiv:2006.06676*, 2020.

[15] Satoshi Suzuki et al., "Topological structural analysis of digitized binary images by border following," *Computer vision, graphics, and image processing*, vol. 30, no. 1, pp. 32–46, 1985.

[16] Saining Xie and Zhuowen Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[17] 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," *arXiv preprint arXiv:1706.08500*, 2017.

[18] 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.

[19] Xun Huang and Serge Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[20] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, "Diverse image-to-image translation via disentangled representations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 35–51.