Language-Driven Artistic Style Transfer

Tsu-Jui Fu[†], Xin Eric Wang[‡], William Yang Wang[†]

[†]UC Santa Barbara [‡]UC Santa Cruz
{tsu-juifu, william}@cs.ucsb.edu xwang366@ucsc.edu

A Complete Results on all Baselines

Table 1 shows the complete LDAST results using visual attribute instructions on DTD^2 [2]. As with previous style transfer methods, NST [5] and WCT [13] cannot handle the target style well (higher Percept loss), resulting in the irrelevant produced patterns to style instructions (lower VLS). AdaIn [7] better performs stylization from guided texts, but the content will be much modified with a relatively lower SSIM. Similar observations can be found in Fig. 1, where the scissors and color pencils by AdaIn are mostly distorted in the second row. In contrast, CLVA presents the detailed multi-color pattern (the second case) yet preserves the concrete structure of the flower (the third case) at the same time.

We also illustrate the results of SANet [15] and LST [12] on specific content domain (Car and Church) in Table 2. By learning from pairs of style instructions and images, style transfer methods can perform better stylization than StyleGAN-restricted methods (StyleCLIP [16] and NADA [4]). Despite having similar results on automatic metrics, there is a noticeable quality gap compared to our CLVA in Fig. 2. Unlike SANet and LST, which contain repetitive and chaotic patterns, CLVA accomplishes LDAST with even more concrete contents (the second car and the fourth church) than Semi-GT through consistent matching during contrastive reasoning.

Method	Automatic Metrics			
	SSIM↑	Percept↓	FAD↓	VLS↑
NST [5]	19.70	0.2441	0.1920	18.97
WCT [13]	**37.88**	0.2617	0.1720	19.76
AdaIn [7]	29.43	<u>0.2081</u>	0.1748	21.65
SANet [15]	35.50	0.2129	0.1627	23.57
LST [12]	34.84	0.2137	<u>0.1533</u>	23.16
ManiGAN [11]	32.7	0.2401	0.1663	23.25
CLIPstyler [10]	25.24	0.2598	0.1818	**24.62**
CLVA	<u>36.65</u>	**0.2033**	**0.1493**	<u>24.00</u>

Table 1: Complete results using visual attribute instructions on DTD^2.

Method	Automatic Metrics			
	SSIM↑	Percept↓	FAD↓	VLS↑
SANet [15]	30.95	<u>0.1982</u>	0.1638	23.20
LST [12]	**31.16**	0.2045	<u>0.1606</u>	23.34
ManiGAN [11]	26.45	0.2329	0.1672	23.44
StyleCLIP [16]	28.03	0.2609	0.1812	21.55
NADA [4]	16.98	0.2733	0.1876	23.38
CLIPstyler [10]	18.43	0.2493	0.1826	**24.16**
CLVA	<u>30.98</u>	**0.1957**	**0.1544**	<u>23.68</u>

Table 2: Complete results on specific content domain (Car and Church).

B Implementation Detail

We adopt VGG-19 [21,15] as our visual encoder G_{E} and visual decoder G_{D}. Text encoder ϕ first adopts RoBERTa [14,17] for a general linguistic and then

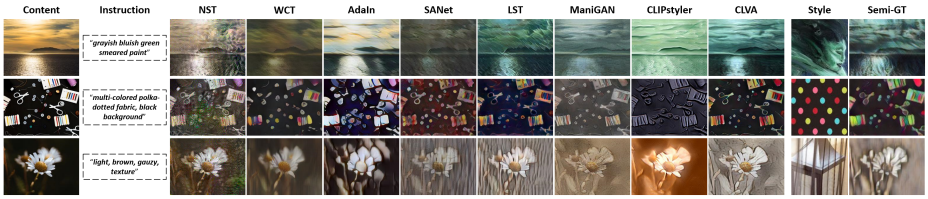

Fig. 1: Complete visualization using visual attribute instructions on DTD[2].

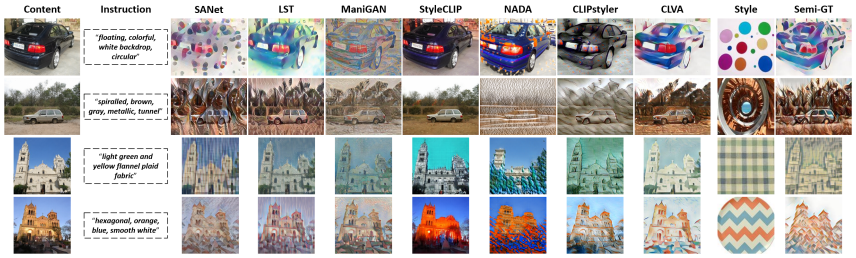

Fig. 2: Complete visualization on specific content domain (Car and Church).

expands its spatial dimension to jointly embed with style features. We follow the self-attention layer from SANet [15] to fuse between content and style features in G_{D}. The patch-wise style discriminator D contains a similar architecture with a dense layer to determine the correlation between instructions and image patches. Both G_{E} and G_{D} are initialized from SANet and further update during the CLVA training process. We adopt Adam [9] to optimize CLVA with learning rate 3e-4 for $\mathcal{L}_G$, 1e-4 for $\mathcal{L}_D$, and 3e-5 for $\mathcal{L}_{\mathrm{ctr}}$.

C Retrieval-based Baseline

Apart from producing the transferred result by the style instruction $\mathcal{X}$, we also investigate a two-step retrieval-based baseline. We first adopt the CLIP [20] alignment to find the most similar style image $\mathcal{S}$ (with 13.9 R@1 and 30.7 R@5 on DTD[2] [2]) via $\mathcal{X}$. Then, the retrieved $\mathcal{S}$ is used to carry out standard style transfer. Table 3 shows that the two-step retrieval baseline performs slightly better on visual similarity to Semi-GT. However, by stylization from guided texts, CLVA produces more correlated style patterns to instructions (higher 24.00 VLS). In addition, this retrieval-based method still relies on an existing set of style images and limits the diversity of stylization due to the collection size.

	Automatic Metrics			
Method	SSIM↑	Percept↓	FAD↓	VLS↑
SANet [15]	35.50	0.2129	0.1627	23.57
SANet (rtrv.)	**37.74**	**0.2005**	**0.1421**	<u>23.68</u>
CLVA	<u>36.65</u>	<u>0.2033</u>	<u>0.1493</u>	**24.00**

Table 3: Testing results of the two-step retrieval-based baseline using visual attribute instructions on DTD[2].

D Human Evaluation

We investigate the quality of LDAST results from the human aspect through Amazon Mechanical Turk. Fig. 3 illustrates the screenshots of the human tasks. MTurkers rank the correlation of the LDAST result according to Content, Instruction, Style, and Semi-GT matching. Each MTurker rewards $2.0 and takes a mean of 15 minutes.

Fig. 3: The screenshots of the ranking tasks for human evaluation on LDAST.

E Limitation and Ethics Discussion

Though our work benefits creative visual applications, there are several remaining technical issues. At first, complicated instructions that contain excessive visual attributes or emotional effects are still difficult to address by our CLVA. CLVA may lean towards specific visual concepts, resulting in correlated but monotonous stylization. Secondly, since the learning of CLVA relies on patch-wise style discriminator D, the quality of the randomly sampled patches will crucially influence the transferred results. On the other hand, there may be a "fake as real" doubt for those manipulated content images. To mitigate this issue, we can apply techniques from image forensics [22,8,3] to detect the authenticity of an image. Regarding guided instructions, for example, hate speech detection [1,6,19,18] can help to filter out malicious texts and prevent from producing controversial results with ethics concerns.

References

1. Aluru, S.S., Mathew, B., Saha, P., Mukherjee, A.: Deep Learning Models for Multilingual Hate Speech Detection. In: ECML-PKDD (2020)

2. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing Textures in the Wild. In: CVPR (2014)
3. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging Frequency Analysis for Deep Fake Image Recognition. In: ICML (2020)
4. Gal, R., Patashnik, O., Maron, H., Chechik, G., Cohen-Or, D.: StyleGAN-NADA: CLIP-Guided Domain Adaptation of Image Generators. In: arXiv:2108.00946 (2021)
5. Gatys, L.A., Ecker, A.S., Bethge, M.: A Neural Algorithm of Artistic Style. In: arXiv:1508.06576 (2015)
6. Huang, X., Xing, L., Dernoncourt, F., Paul, M.J.: Multilingual Twitter Corpus and Baselines for Evaluating Demographic Bias in Hate Speech Recognition. In: LREC (2020)
7. Huang, X., Belongie, S.: Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. In: ICCV (2017)
8. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting Fake News: Image Splice Detection via Learned Self-Consistency. In: ECCV (2018)
9. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR (2015)
10. Kwon, G., Ye, J.C.: CLIPstyler: Image Style Transfer with a Single Text Condition. In: CVPR (2022)
11. Li, B., Qi, X., Lukasiewicz, T., Torr, P.H.S.: ManiGAN: Text-Guided Image Manipulation. In: CVPR (2020)
12. Li, X., Liu, S., Kautz, J., Yang, M.H.: Learning Linear Transformations for Fast Arbitrary Style Transfer. In: CVPR (2019)
13. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., Yang, M.H.: Universal Style Transfer via Feature Transforms. In: NeurIPS (2017)
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. In: arXiv:1907.11692 (2019)
15. Park, D.Y., Lee, K.H.: Arbitrary Style Transfer with Style-Attentional Networks. In: CVPR (2019)
16. Patashnik, O., Wu, Z., Shechtman, E., Cohen-Or, D., Lischinski, D.: StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In: ICCV (2021)
17. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: EMNLP (2019)
18. Samanta, B., Ganguly, N., Chakrabarti, S.: Improved Sentiment Detection via Label Transfer from Monolingual to Synthetic Code-Switched Text. In: ACL (2019)
19. Samghabadi, N.S., Patwa, P., PYKL, S., Mukherjee, P., Das, A., Solorio, T.: Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In: LREC (2020)
20. Shi, L., Shuang, K., Geng, S., Su, P., Jiang, Z., Gao, P., Fu, Z., de Melo, G., Su, S.: Contrastive Visual-Linguistic Pretraining. In: arXiv:2007.13135 (2020)
21. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. In: ICLR (2015)
22. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-Generated Images are Surprisingly Easy to Spot...for Now. In: CVPR (2020)