SCSA: A Plug-and-Play Semantic Continuous-Sparse Attention for Arbitrary Semantic Style Transfer

Supplementary Material

A. Societal Impact

To achieve a thorough understanding of the broader implications of our study, we assess the proposed method in terms of both positive and negative impacts, followed by suggestions to mitigate any adverse effects.

Positive Impact. Our proposed method offers several societal benefits across different sectors. (1) This method provides researchers with innovative tools, encouraging them to explore challenges in semantic style transfer with fresh perspectives. By promoting the development of simple and effective solutions, our approach plays a role in advancing scientific discovery and accelerating innovation. (2) Artists can leverage this method to enhance their creative workflow and optimize productivity. The tool not only allows them to explore new artistic possibilities but also supports them in creating high-quality works more efficiently. Additionally, for general users, this method offers an accessible means to produce visually appealing outputs that align with semantic needs, enriching their creative experience. (3) By democratizing artistic tools and lowering barriers to creating stylistically unique digital art, this method encourages wider participation in creative expression. Such accessibility can lead to increased cultural exchange, as individuals from diverse backgrounds share and interpret art through new stylistic forms. This ultimately strengthens social cohesion and broadens avenues for cultural representation.

Negative Impact. While beneficial, the method also presents potential risks that warrant attention. (1) As more people adopt AI-assisted tools, there may be a gradual erosion of unique artistic expression, with creators becoming reliant on automated processes instead of traditional, personal techniques. This could lead to a homogenization of digital art, where distinct styles and individual creativity are diminished. (2) In the wrong hands, this technology could be utilized to produce manipulated images or videos that appear authentic, fostering misinformation. Misuse of stylization techniques in this manner poses ethical concerns and risks misleading audiences if safeguards are not established.

Mitigation Strategies. To counteract these negative impacts, we propose the following measures: (1) By supporting customization features and encouraging users to incorporate personal touches in AI-assisted creations, we aim to preserve artistic diversity and prevent homogenization. Enabling flexibility in the method's application allows artists to maintain their unique creative identities while benefiting from the technology. (2) Educating users on the ethical implications and potential misuse of stylization techniques is vital. By fostering an understanding of responsible AI use, we can mitigate the risk of harmful applications and promote ethical practices.

B. Used Assets

We utilize the following assets for our experiments:

- SANet [15]: https://github.com/GlebSBrykin/SANET, MIT license.
- StyTr^2 [6]: https://github.com/diyiiyiii/StyTR-2, No License.
- StyleID [4]: https://github.com/jiwoogit/StyleID, MIT license.
- STROTSS [11]:https://github.com/nkolkin13/STROTSS, No License.
- MAST [9]:https://github.com/NJUHuoJing/MAST, No License.
- TR [24]:https://github.com/EndyWon/Texture-Reformer, MIT license.
- DIA [13]:https://github.com/harveyslash/Deep-Image-Analogy-PyTorch, MIT license.
- GLStyleNet [22]:https://github.com/EndyWon/GLStyleNet, MIT license.

To the best of our knowledge, there are no moral or ethical concerns associated with these assets. We have thoroughly reviewed their use to ensure compliance with ethical standards, confirming that their implementation aligns with responsible research practices. This careful consideration reinforces our commitment to conducting research that adheres to both ethical guidelines and scientific integrity.

C. Application Details

We select three representative Attn-AST methods for SCSA embedding experiments: SANet [15], built on the CNN framework, StyTR2 [6], based on the Transformer architecture, and StyleID [4], utilizing the Diffusion model. By integrating SCSA into these Attn-AST methods, we aim to investigate its performance in semantic style transfer tasks.

C.1. SANet with SCSA

Given a quadruple data $\{I_c, I_{csem}, I_s, I_{ssem}\}$, consisting of a content image I_c and its semantic map I_{csem}, a style image I_s and its semantic map I_{ssem}, and the SANet model $M = \{E, T_{UA}, D\}$, which is composed of an encoder E, a feature transformation module T_{UA}, and a decoder D, we

aim to generate a stylized image I_{cs} that meets the semantic needs by replacing T_{UA} with our T_{SCSA}.

In the initial stage, we get the encoded quadruple features $\{F_c, F_{csem}, F_s, F_{ssem}\}$:

$$\begin{aligned} F_c = E(I_c),\ F_s = E(I_s), \\ F_{csem} = E(I_{csem}),\ F_{sem} = E(I_{sem}), \end{aligned} \tag{1}$$

where E is a VGG-19 [18] encoder.

Subsequently, we feed the quadruple features into our feature transformation module T_{SCSA}. Thus, the following operations will be performed:

$$\begin{aligned} Q_1 = f_q(\bar{F}_{csem}),\ K_1 = f_k(\bar{F}_{ssem}),\ V_1 = f_v(F_s), \\ \bar{\mathcal{A}} = G_1(Q_1^{\mathsf{T}} \otimes K_1), \\ F_{sca} = f_o(softmax(\bar{\mathcal{A}}) \otimes V_1), \end{aligned} \tag{2}$$

where $\bar{F}_x$ represents the normalized form of the features F_x using its mean and standard deviation. f_q, f_k, f_v, and f_o are the projection networks in T_{UA}. G_1 is the modulation function in Eq.6 of the main paper. F_{sca} are the features that contain the overall style characteristics of the corresponding semantic regions. Then, we obtain the new encoded content features using semantic adaptive instance normalization in Sec.3.1 of the main paper:

$$F_c = \ S\text{-}AdaIN(F_c, F_s), \tag{3}$$

where the new content features F_c partially eliminate the influence of the inherent style of the original encoded features of the content image, thereby offering a more precise query for subsequent SSA and content features with distinct stylistic characteristics of the style features. Hence, the following formulas exist for SSA:

$$\begin{aligned} Q_2 = f_q(\bar{F}_c),\ K_2 = f_k(\bar{F}_s),\ V_2 = f_v(F_s), \\ \bar{\mathcal{B}} = G_2(Q_2^{\mathsf{T}} \otimes K_2), \\ F_{ssa} = f_o(softmax(\bar{\mathcal{B}}) \otimes V_2), \end{aligned} \tag{4}$$

where $\bar{F}_x$, f_q, f_k, f_v, and f_o are the same as those in Eq. 2. G_2 is the modulation function in Eq.11 of the main paper. F_{ssa} are the features characterized by fine and specific style textures in the corresponding semantic regions. Then, the stylized features through the feature transformation module T_{SCSA} can be obtained:

$$F_{cs} = \alpha_1 \times F_{sca} + \alpha_2 \times F_{ssa} + F_c, \tag{5}$$

where α_1 and α_2 indicate separately the stylization degree for the overall style and vivid textures in semantic regions. We set $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$ in the main paper.

Ultimately, the stylized image can be produced:

$$I_{cs} = D(F_{cs}), \tag{6}$$

where the decoder D structure mirrors that of VGG-19.

C.2. StyTr^2 with SCSA

Given a quadruple data $\{I_c, I_{csem}, I_s, I_{ssem}\}$, consisting of a content image I_c and its semantic map I_{csem}, a style image I_s and its semantic map I_{ssem}, and the StyTr^2 model $M = \{E_c, E_s, T_{UA}, D\}$, which is composed of a content encoder E_c, a style encoder E_s, some feature transformation transformer modules T_{UA}, and a decoder D, we aim to generate a stylized image I_{cs} that meets the semantic needs by replacing T_{UA} with our T_{SCSA}.

As a first step, we get the encoded quintuple features $\{F_c, F_{csem}, F_s^c, F_s^s, F_{ssem}\}$:

$$\begin{aligned} F_c = E_c(I_c),\ F_s^c = E_c(I_s),\ F_s^s = E_s(I_s), \\ F_{csem} = E_c(I_{csem}),\ F_{sem} = E_s(I_{sem}), \end{aligned} \tag{7}$$

where E_c and E_s are the content and style transformer [21] encoders, respectively.

Consequently, we feed the quintuple features into our feature transformation module T_{SCSA}. Thus, the following operations will be performed:

$$\begin{aligned} Q_1 = f_q(F_{csem}),\ K_1 = f_k(F_{ssem}),\ V_1 = f_v(F_s^s), \\ \bar{\mathcal{A}} = G_1(Q_1^{\mathsf{T}} \otimes K_1), \\ F_{sca} = f_o(softmax(\bar{\mathcal{A}}) \otimes V_1), \end{aligned} \tag{8}$$

where f_q, f_k, f_v, and f_o are the projection networks in T_{UA}. G_1 is the modulation function in Eq.6 of the main paper. F_{sca} are the features that contain the overall style characteristics of the corresponding semantic regions. Then, we obtain the new encoded content features using semantic adaptive instance normalization in Sec.3.1 of the main paper:

$$F_c^{S\text{-}AdaIN} = \ S\text{-}AdaIN(F_c, F_s^c), \tag{9}$$

where the new content features $F_c^{S\text{-}AdaIN}$ partially eliminate the influence of the inherent style of the original encoded features of the content image, thereby offering a more precise query for subsequent SSA and content features with distinct stylistic characteristics of the style features. Hence, the following formulas exist for SSA:

$$\begin{aligned} Q_2 = f_q(F_c^{S\text{-}AdaIN}),\ K_2 = f_k(F_s^s),\ V_2 = f_v(F_s^s), \\ \bar{\mathcal{B}} = G_2(Q_2^{\mathsf{T}} \otimes K_2), \\ F_{ssa} = f_o(softmax(\bar{\mathcal{B}}) \otimes V_2), \end{aligned} \tag{10}$$

where f_q, f_k, f_v, and f_o are the same as those in Eq. 8. G_2 is the modulation function in Eq.11 of the main paper. F_{ssa} are the features characterized by fine and specific style textures in the corresponding semantic regions. Then, the stylized features through the feature transformation module T_{SCSA} can be obtained:

$$F_{cs} = \alpha_1 \times F_{sca} + \alpha_2 \times F_{ssa} + b \times F_c^{S\text{-}AdaIN} + (1-b) \times F_c, \tag{11}$$

where α_1 and α_2 indicate separately the stylization degree for the overall style and vivid textures in semantic regions. We set $\alpha_1 = 1.2$ and $\alpha_2 = 0.5$. b is used to trade off the degree of the semantic style initialization of the content features and the degree of content preservation. We set $b = 0.7$. F_{cs} are utilized as the new content features for the subsequent feature transformation modules. It is important to note that we focus on aligning the semantic style of content features only in the first feature transformation module, while $b = 0$ is set for the remaining feature transformation modules.

Ultimately, the stylized image can be produced:

$$I_{cs} = D(F_{cs}), \tag{12}$$

where the even-numbered transformers of the decoder D are replaced with our T_{SCSA}.

C.3. StyleID with SCSA

Given a quadruple data $\{I_c, I_{csem}, I_s, I_{ssem}\}$, consisting of a content image I_c and its semantic map I_{csem}, a style image I_s and its semantic map I_{ssem}, and the StyleID model $M = \{E, U\text{-}Net, D\}$, which is composed of a encoder E, a denoising model $U\text{-}Net$ [17], and a decoder D, we aim to generate a stylized image I_{cs} that meets the semantic needs by replacing certain T_{UA} of $U\text{-}Net$ module with T_{SCSA}.

First of all, we obtain the encoded quadruple features $\{X_{c0}, X_{csem0}, X_{s0}, X_{ssem0}\}$ at the time step $t = 0$:

$$\begin{aligned} X_{c0} = E(I_c),\ X_{s0} = E(I_s),\\ X_{csem0} = E(I_{csem}),\ X_{sem0} = E(I_{sem}), \end{aligned} \tag{13}$$

where E is a VAE [10] encoder.

Following that, we acquire the noisy quadruple features $\{F_{cT}, F_{csemT}, F_{sT}, F_{ssemT}\}$ from certain T_{UA} layers of $U\text{-}Net$ at the time step $t = T$ via DDIM inversion [19]:

$$\begin{aligned} F_{cT} = DD\text{-}SA(X_{c0}),\ F_{sT} = DD\text{-}SA(X_{s0}),\\ F_{csemT} = DD\text{-}SA(X_{csem0}),\ F_{semT} = DD\text{-}SA(X_{sem0}), \end{aligned} \tag{14}$$

where $DD\text{-}SA$ represents the DDIM inversion and extracting features from the T_{UA} layers.

As a next step, we obtain the new noisy content features using semantic adaptive instance normalization in Sec.3.1 of the main paper:

$$F_{cT} = S\text{-}AdaIN(F_{cT}, F_{sT}), \tag{15}$$

where the new features F_{cT} partially eliminate the influence of the inherent style of the original noisy features. This provides a more precise query for the subsequent SCA and SSA, along with content features that exhibit distinct stylistic characteristics from the style features.

Following this, we feed the noisy quadruple features into our feature transformation module T_{SCSA}. Thus, the following operations will be performed:

$$\begin{aligned} Q_1 = f_q(t_1 \times \bar{F}_{csemT} + (1 - t_1) \times \bar{F}_{cT}),\\ K_1 = f_k(t_1 \times \bar{F}_{ssemT} + (1 - t_1) \times \bar{F}_{sT}),\\ V_1 = f_v(F_{sT}),\\ \bar{\mathcal{A}} = G_1(Q_1^{\mathsf{T}} \otimes K_1),\\ F_{sca} = f_o(softmax(\bar{\mathcal{A}}) \otimes V_1), \end{aligned} \tag{16}$$

where $\bar{F}_x$ represents the normalized form of the features F_x using its normalized networks in T_{UA}. t_1 represents the trade-off between semantic stylization and content preservation. We set $t_1 = 0.3$. f_q, f_k, f_v, and f_o are the projection networks in T_{UA}. G_1 is the modulation function in Eq.6 of the main paper. F_{sca} are the features that contain the overall style characteristics of the corresponding semantic regions. Similarly, the following formulas exist for SSA:

$$\begin{aligned} Q_2 = f_q(t_2 \times \bar{F}_{cT} + (1 - t_2) \times \tilde{F}_T),\\ K_2 = f_k(\bar{F}_{sT}),\ V_2 = f_v(F_{sT}),\\ \bar{\mathcal{B}} = G_2(Q_2^{\mathsf{T}} \otimes K_2),\\ F_{ssa} = f_o(softmax(\bar{\mathcal{B}}) \otimes V_2), \end{aligned} \tag{17}$$

where $\bar{F}_x$, f_q, f_k, f_v, and f_o are the same as those in Eq. 14. F_T represent the input features of T_{UA}. $\tilde{F}_T$ represents the normalized form of the features F_T using its mean and standard deviation. t_2 represents the trade-off between content preservation and semantic stylization, similar to that in [4]. We set $t_2 = 0.5$. G_2 is the modulation function in Eq.11 of the main paper. F_{ssa} are the features characterized by fine and specific style textures in the corresponding semantic regions. Then, the stylized features through the feature transformation module T_{SCSA} can be obtained:

$$F_{cs} = \alpha_1 \times F_{sca} + \alpha_2 \times F_{ssa} + F_{cT}, \tag{18}$$

where α_1 and α_2 indicate separately the stylization degree for the overall style and vivid textures in semantic regions. We set $\alpha_1 = 0.8$ and $\alpha_2 = 0.2$ in the main paper. It is worth noting that we use the features F_{cT} processed by semantic adaptive instance normalization only at the time step $t = T$, while replacing it with the input features of T_{UA} at other time steps.

Ultimately, the stylized image can be produced:

$$I_{cs} = D(X_{cs0}), \tag{19}$$

where X_{cs0} represents the output of the $U\text{-}Net$ at the time step $t = 0$ through the above DDIM sample. D is a VAE [10] decoder.

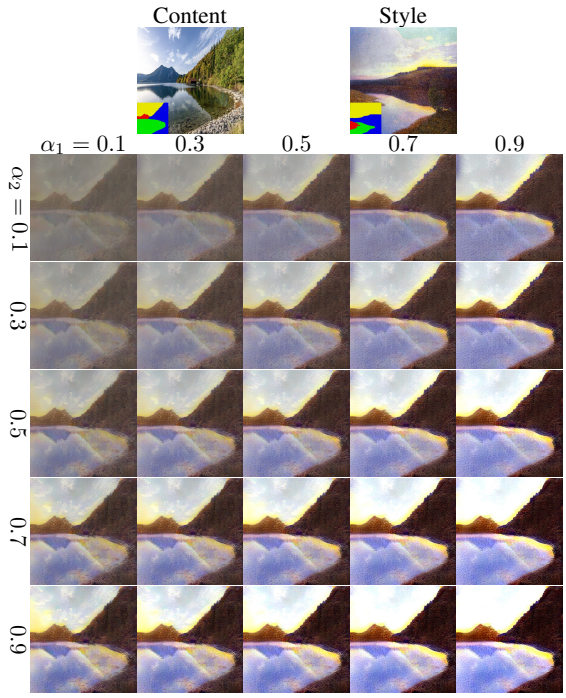


Figure 1. Comparisons of the overall style and local texture intensity in semantic regions of SANet embedded with our SCSA.

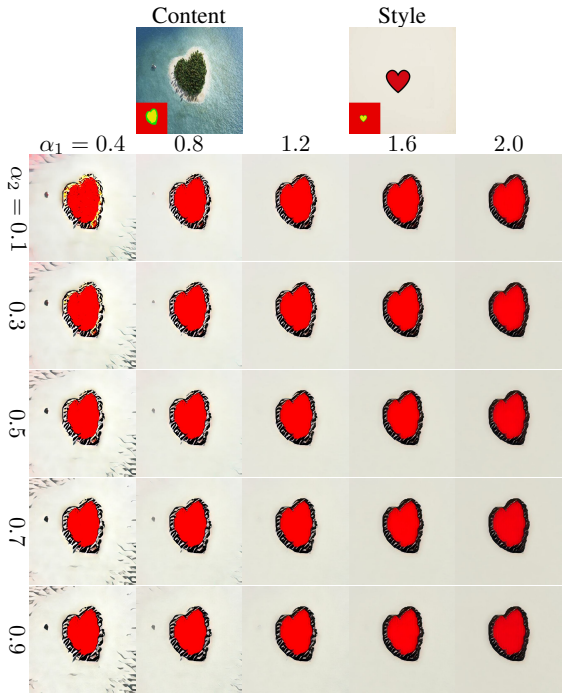


Figure 2. Comparisons of the overall style and local texture intensity in semantic regions of StyTr^2 embedded with our SCSA.

D. Parameter Analysis

To achieve a more comprehensive understanding of the effects of experimental parameters, we also perform additional experiments that delve into their specific roles and contributions, providing deeper insights into the effectiveness of our method across various conditions.

Overall Style-Local Texture Intensity. Our SCSA enables dynamic adjustment of the intensity of overall style and local textures in corresponding semantic regions. To demonstrate this, we conduct relevant experiments for comprehensive analyses.

As shown in Fig. 1, Fig 2, and Fig 3 an increase in the parameter α_1 enhances the overall style expression of the semantic regions. In contrast, a rise in parameter α_2 brings greater clarity to the textures within these regions.

Content-Style Trade-off. Through adjustments to parameter b in Sec. C.2 and parameters t_1 and t_2 in Sec. C.3, our method offers a flexible balance between stylization intensity and content preservation. To illustrate this, we perform a series of targeted experiments, allowing for an in-depth analysis and comprehensive evaluation. It is important to highlight that the effectiveness of t_2 has already been established in [4]. Thus, our focus here will be solely on demonstrating the content preservation capability of t_1, thoroughly examining its impact on maintaining the integrity of the original content during the stylization process.

As illustrated in Fig. 4, with the increase of the parameter b, the degree of stylization progressively intensifies, while the level of content preservation experiences a slight

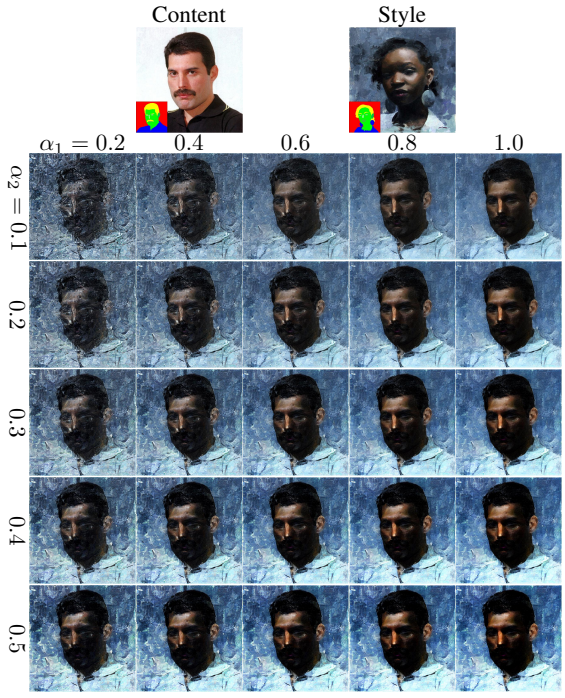


Figure 3. Comparisons of the overall style and local texture intensity in semantic regions of StyleID embedded with our SCSA.

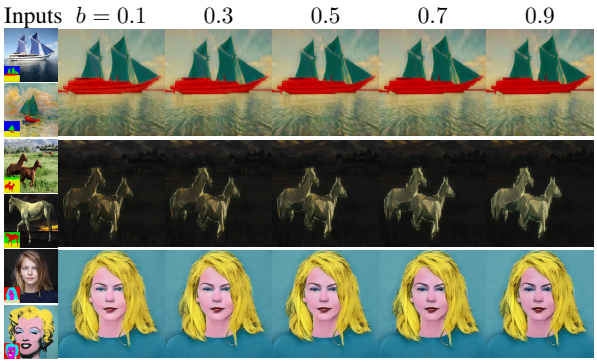


Figure 4. Comparisons of the trade-off between content preservation and stylization of StyTr^2 embedded with SCSA.

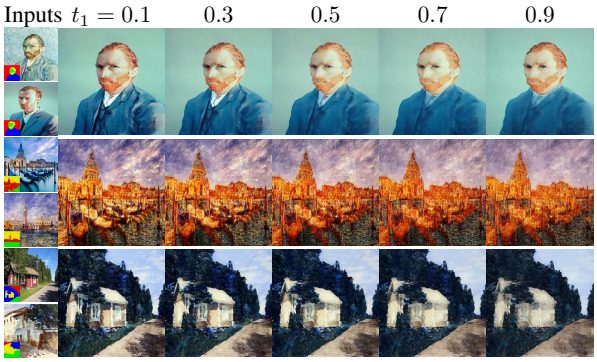


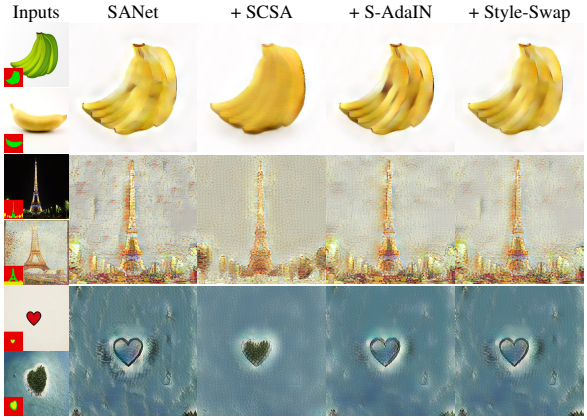
Figure 5. Comparisons of the trade-off between content preservation and stylization of StyleID embedded with SCSA.

decline. This phenomenon is particularly evident across different subjects. For instance, in the $1st$ row, the boat's structure becomes more pronounced in the stylization process; while its basic form remains discernible, the accuracy of content preservation decreases slightly. In the $2nd$ row, the eyes of the horse exhibit more distinctive stylistic features, although their original characteristics are somewhat retained, the details appear less clear. In the $3rd$ row, the eyes of the person maintain their fundamental structure while integrating more stylistic elements, which also affects their level of content preservation.

As shown in Fig. 5, as t_1 increases, the degree of stylization in semantic regions gradually intensifies, while the ability to maintain content declines, e.g., the building structures in the $1st$ row, the cloth in the $2nd$ row the house in the $3rd$ row. Therefore, t_1 allows for a trade-off between semantic stylization and content preservation.

E. Dataset Details

We select content and style images from prior research [1, 2, 8, 12, 20, 23, 25–29], publicly available datasets [5, 7, 16, 30], and the Internet to ensure a diverse and comprehensive set of images. These sources provide a broad spectrum of content and style representations, encompassing various domains and visual styles. We then construct the semantic maps for each image, capturing their intrinsic features and structural elements. Based on these maps, we generate 91 validated quadruple data, which will be made publicly available upon acceptance of the paper to foster further research and development in semantic style transfer.



SSL ↓	1.6583	**0.8762**	1.7289	1.5751
FID ↓	14.3385	**13.0788**	14.2748	14.1523
CFSD ↓	0.1103	**0.0874**	0.1057	0.1077

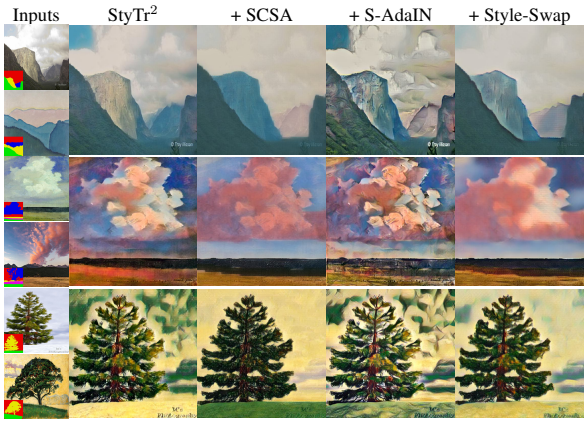
Figure 6. Qualitative and quantitative comparisons among CNN-based SANet, SANet with SCSA, SANet with S-AdaIN, and SANet with Style-Swap.

F. Ablation Study

To comprehensively validate the superiority of our proposed SCSA method, we incorporate the existing S-AdaIN [14] and Style-Swap [3] techniques into the universal attention module of the Attn-AST framework.

SCSA vs. S-AdaIN. As shown in Fig. 6, Fig. 7, and Fig 8, S-AdaIN can achieve semantic style transfer to some extent, but the degree of semantic stylization is far inferior to ours. For example, in the $3rd$ row of Fig. 6, our heart features dense tree-like textures, while S-AdaIN lacks these, and its global style of semantic regions is not effectively transferred. In the $1st$ row of Fig. 7, the overall style of our mountains is continuous, whereas S-AdaIN lacks this. A similar difference is in the $2nd$ row in Fig. 8. In addition to the qualitative comparison, the quantitative results from the three figures further demonstrate that the degree of stylization in stylized images generated by S-AdaIN is notably lower than that of our method, as reflected in the higher SSL and FID values. Also, incorporating S-AdaIN into SANet and StyTr^2 leads to inferior content preservation compared to our method, as indicated by higher CFSD values.

SCSA vs. Style-Swap. As shown in Fig. 6, Fig. 7, and Fig. 8, although Style-Swap can achieve some degree of se-



SSL ↓	1.9826	**1.2228**	1.7318	1.3824
FID ↓	12.5273	**12.3963**	14.9473	13.9573
CFSD ↓	0.0752	**0.0705**	0.1007	0.0818

Figure 7. Qualitative and quantitative comparisons among Transformer-based StyTr^2, StyTr^2 with SCSA, StyTr^2 with S-AdaIN, and StyTr^2 with Style-Swap.

SSL ↓	1.7538	**1.2447**	1.6054	2.5032
FID ↓	12.5944	**12.4497**	12.5771	19.3049
CFSD ↓	0.0916	0.1178	0.0939	**0.0732**

Figure 8. Qualitative and quantitative comparisons of Diffusion-based StyleID, StyleID with SCSA, StyleID with S-AdaIN, and StyleID with Style-Swap.

mantic style transfer, it still lacks the accuracy of the transferred semantic style and the continuity within the same semantic region. A case in point is the $3rd$ row of Fig. 6 (the heart), the $3rd$ row of Fig.7 (the grass), and the $1st$ row of Fig. 8 (the background), where the corresponding semantic regions fail to undergo accurate and continuous style transfer. Furthermore, the overall stylization effect of stylized images generated by Style-Swap is significantly inferior to that of our method, as evidenced by its higher SSL and FID values in the three figures compared to SCSA.

The above analysis demonstrates that our SCSA greatly outperforms both S-AdaIN and Style-Swap in qualitative and quantitative perspectives, proving its superiority.

G. User Study

Our user study is primarily divided into two main parts.

In one part of the user study, we presented participants with corresponding data quadruples $\{I_c, I_{csem}, I_s, I_{ssem}\}$ along with a stylized image generated by a traditional Attn-AST method, randomly selected from CNN-based, Transformer-based, and Diffusion-based approaches, and one produced by the corresponding traditional Attn-AST method with our SCSA embedded, both displayed in random order. We asked participants to select, "Which image do you believe represents the most satisfactory result of semantic style transfer?" This part aims to compare user preferences between the SCSA-embedded Attn-AST approaches and the traditional Attn-AST methods, thereby validating the effectiveness and generalization of our SCSA in semantic style transfer.

In another part, we also presented participants with the corresponding data quadruples $\{I_c, I_{csem}, I_s, I_{ssem}\}$, along with three stylized images generated by the traditional Attn-AST method with our SCSA embedded, CNN-based, Transformer-based, and diffusion-based methods, as well as stylized images produced by five SOTA semantic style transfer methods. The display order of eight stylized images was randomized. Again, we asked participants to select the stylized image they found most satisfactory. This part aims to validate the effectiveness of our SCSA in comparison to SOTA semantic style transfer methods, providing a subjective basis for establishing our approach as a new benchmark for semantic style transfer.

We invited 40 participants to take part in our user study, with each participant responding to a total of 30 questions-15 in each of the two previously mentioned sections. This comprehensive survey design enabled us to gather detailed insights into user preferences, ultimately resulting in the collection of 1,200 votes, which will be essential for our analysis and validation of the findings.

H. Additional Experiment

In the main paper, we have verified the generalization of SCSA when integrated in the Attn-AST methods. To further validate the generalization of our proposed SCSA concerning data, we carry out an additional experiment. Specifically, we broaden the scope of semantics, extending them beyond individual instances to encompass regions that share the same semantic mask. In this context, the instance semantics within these regions targeted for style transfer may not be entirely uniform.

As shown in Fig. 9, our SCSA exhibits remarkable data generalization. It is capable of processing not only data with consistent semantic instances but also data with distinct semantic instances, as long as users can provide corresponding identical mask labels.

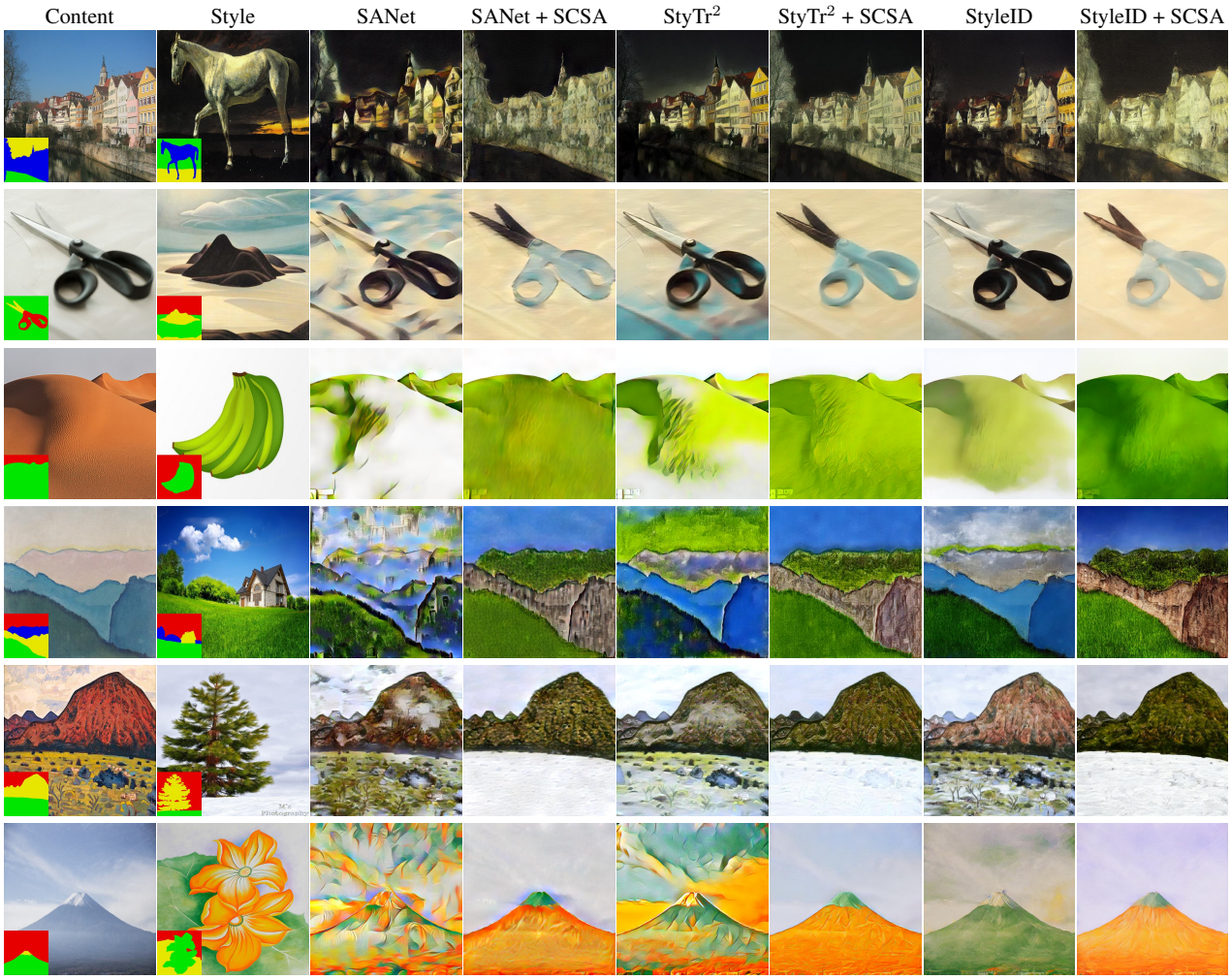


Figure 9. Qualitative comparisons between Attn-AST approaches and they with our SCSA for different semantics.

I. More Discussions

Effects of Attn-AST methods with SCSA. As stated in the main paper, SCSA is a plug-and-play semantic style transfer method. The quality of the stylized images it produces is directly influenced by the capability of the underlying Attn-AST model. Especially, the stronger the transfer style ability of the Attn-AST model, the more remarkable the semantic stylization effects achieved when integrated with SCSA.

Processing costs of Attn-AST methods with SCSA. As SCSA incorporates the processing and guidance of semantics, Attn-AST methods with SCSA require more time and memory than the original Attn-AST. However, our primary focus is currently on the effectiveness of semantic stylization, with efficiency optimization planned as a future research direction.

J. Qualitative Comparison

To conduct a more in-depth evaluation of SCSA, we present a broader array of qualitative results.

As shown in Fig. 10, Fig. 11, Fig. 12, Fig. 13, Fig. 14, Fig. 15, and Fig. 16, our SCSA method seamlessly enhances existing arbitrary style transfer techniques, enabling versatile semantic style transfer with performance that exceeds current state-of-the-art methods in both intensity and stability. Specifically, the stylized images produced by DIA exhibit incomplete contents of the content images. The stylized images generated by TR, STROTSS, and GLStyleNet sometimes incorporate content elements from the style image. MAST occasionally yields semantic stylization results that lack accuracy.

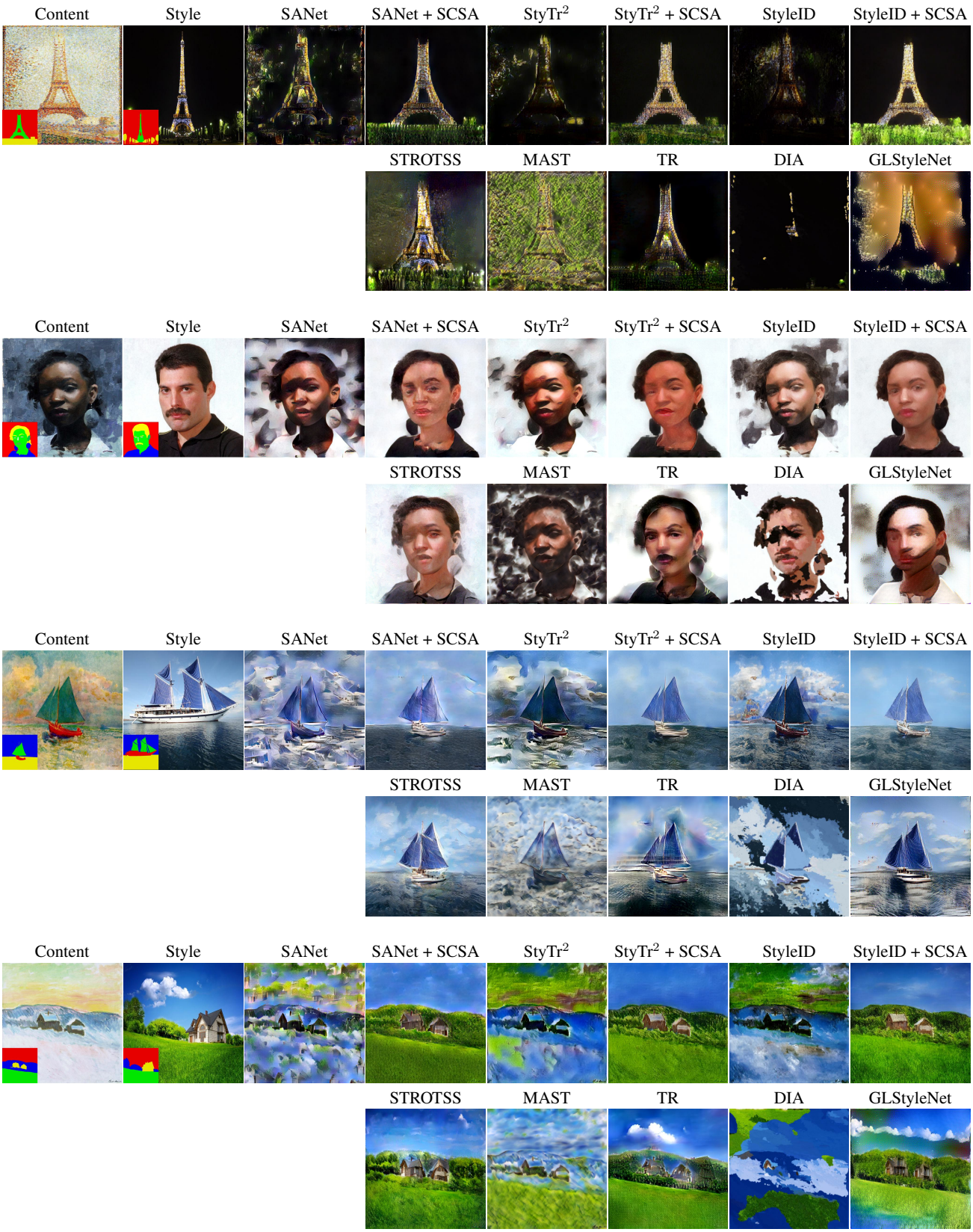


Figure 10. Qualitative comparisons among Attn-AST approaches, those with SCSA, and SOTA methods.

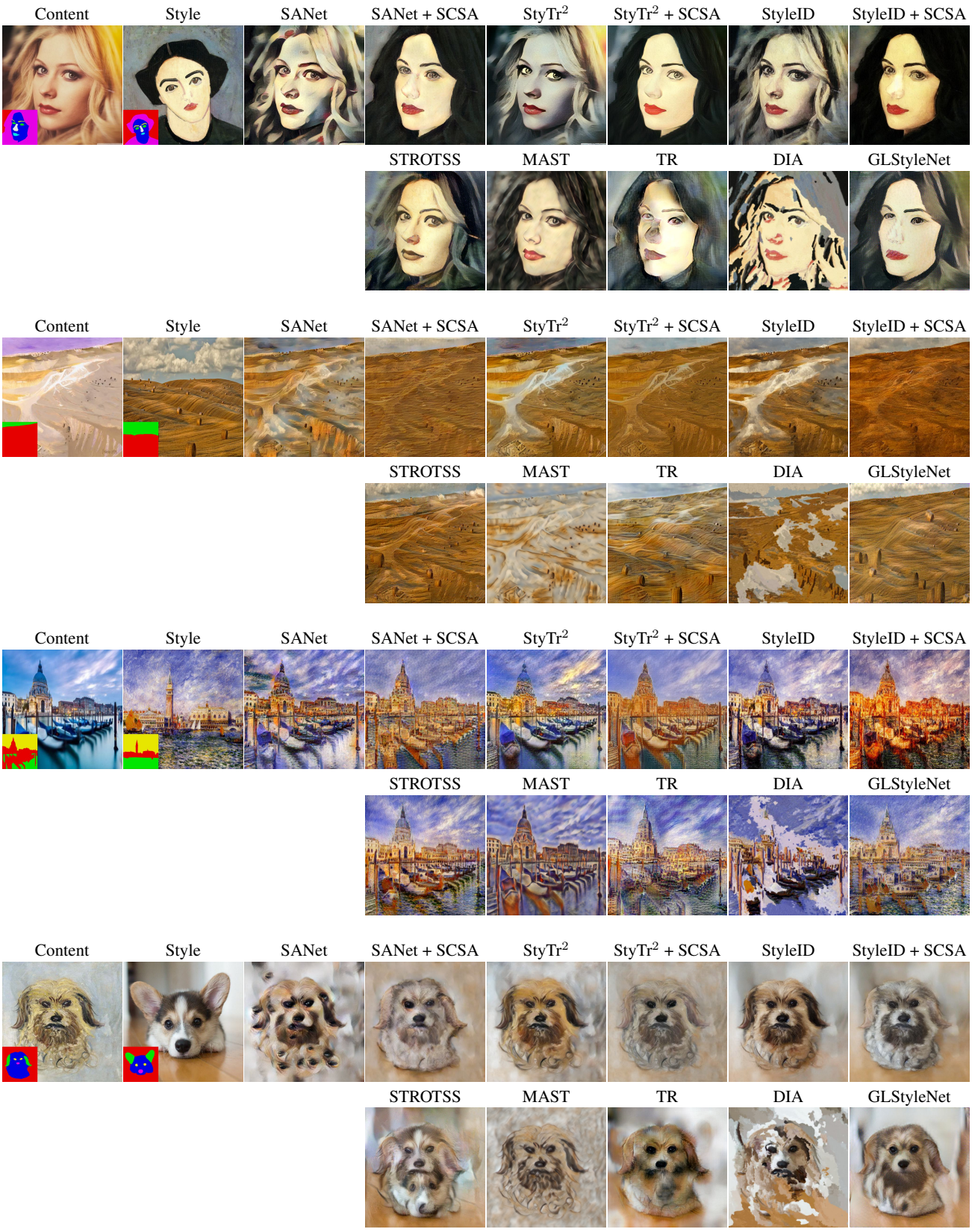


Figure 11. Qualitative comparisons among Attn-AST approaches, those with SCSA, and SOTA methods.

Figure 12. Qualitative comparisons among Attn-AST approaches, those with SCSA, and SOTA methods.

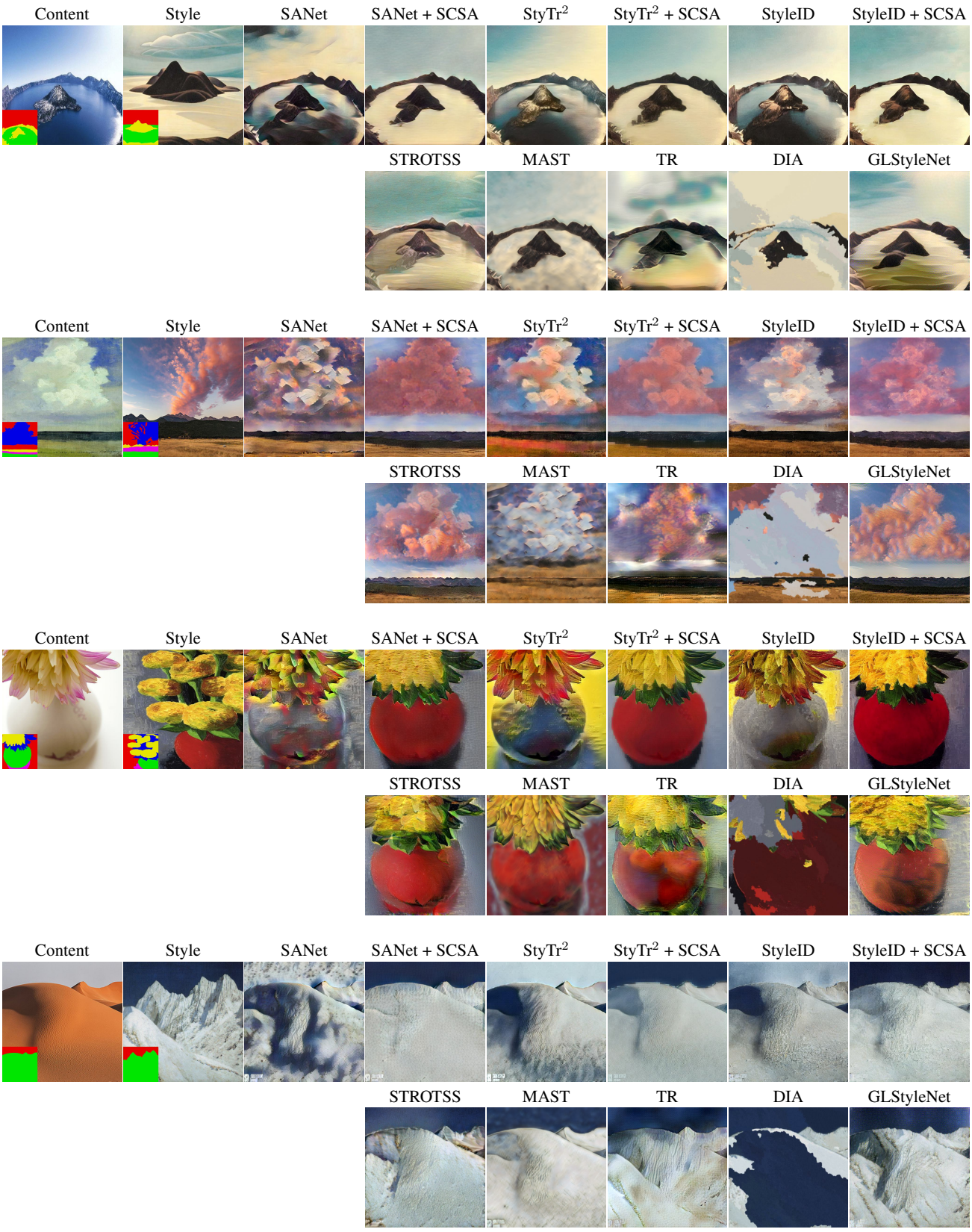


Figure 13. Qualitative comparisons among Attn-AST approaches, those with SCSA, and SOTA methods.

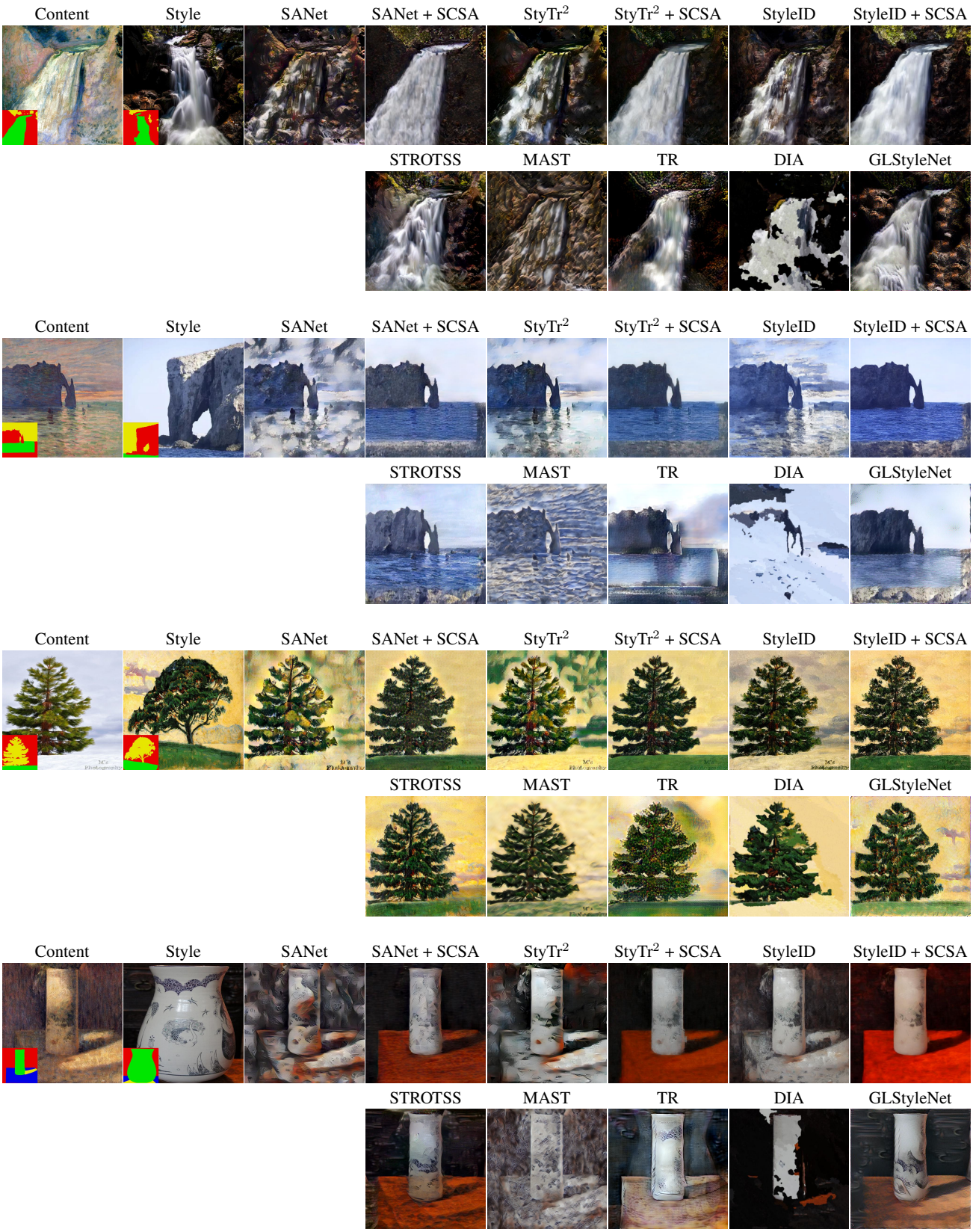


Figure 14. Qualitative comparisons among Attn-AST approaches, those with SCSA, and SOTA methods.

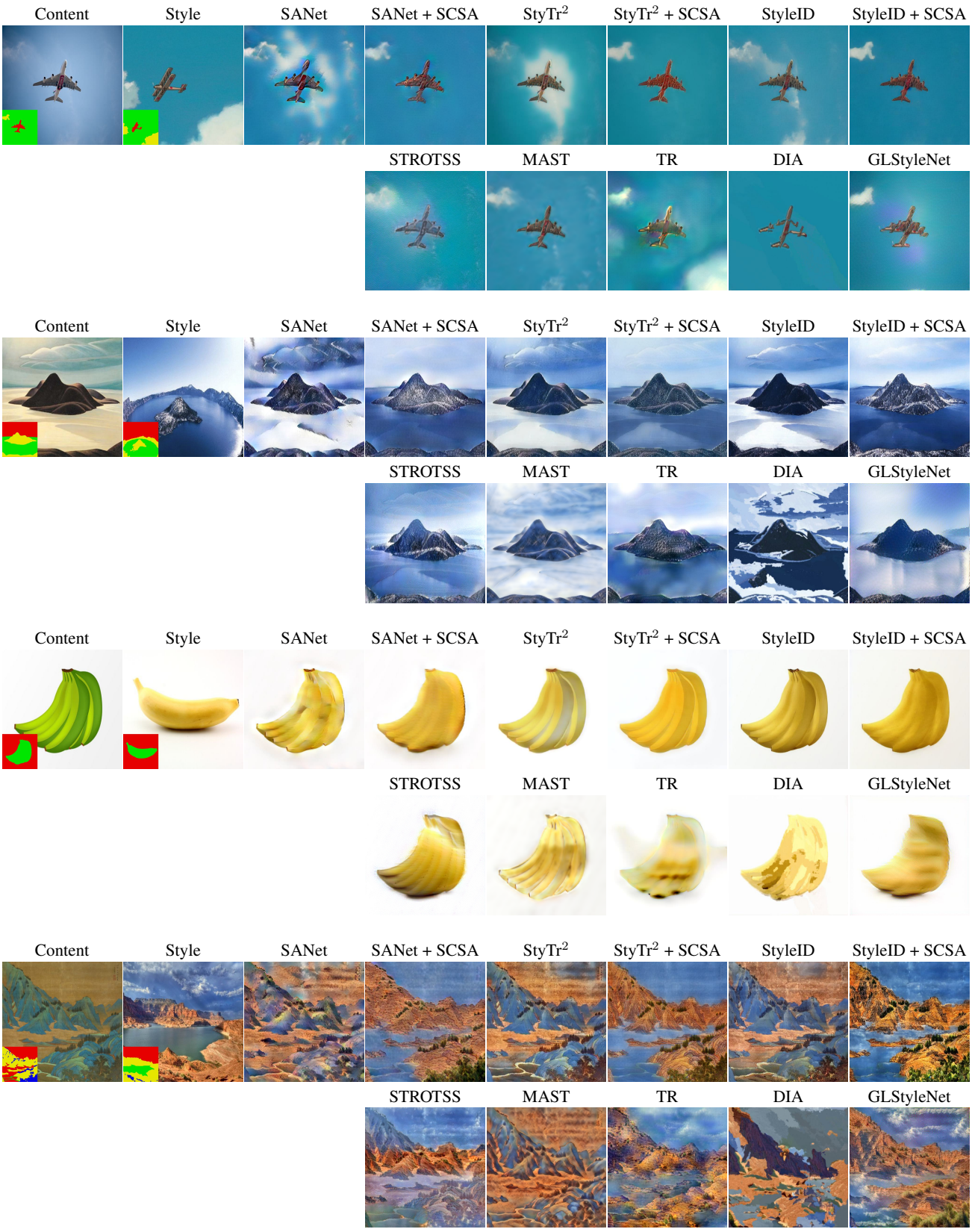


Figure 15. Qualitative comparisons among Attn-AST approaches, those with SCSA, and SOTA methods.

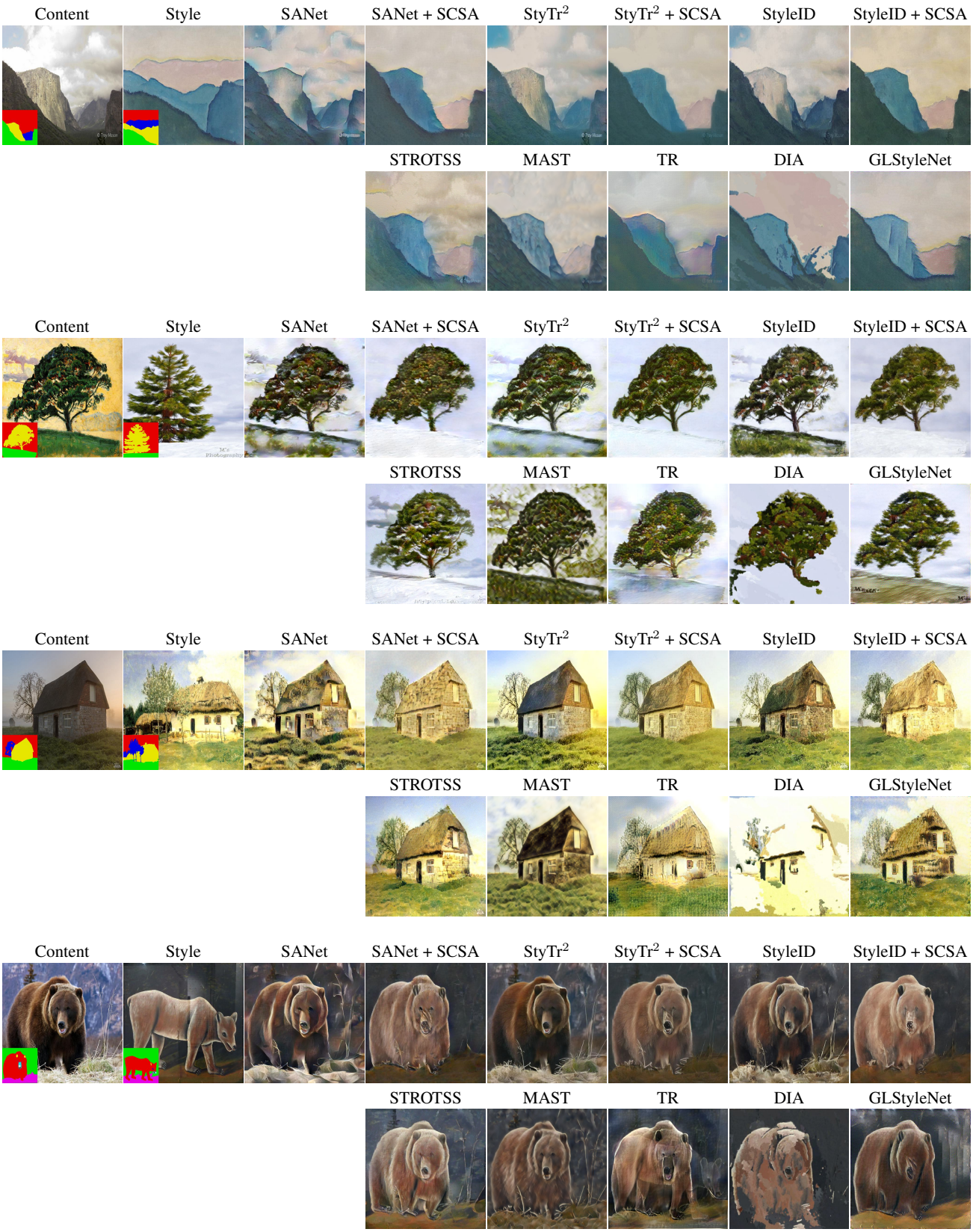


Figure 16. Qualitative comparisons among Attn-AST approaches, those with SCSA, and SOTA methods.

References

[1] Namhyuk Ahn, Junsoo Lee, Chunggi Lee, Kunhee Kim, Daesik Kim, Seung-Hun Nam, and Kibeom Hong. Dreamstyler: Paint by style inversion with text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 674–681, 2024. 5

[2] Alex J Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016. 5

[3] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*, 2016. 5

[4] Jiwoo Chung, Sangeek Hyun, and Jae-Pil Heo. Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8795–8805, 2024. 1, 3, 4

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[6] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022. 1

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5

[8] Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, and Hanspeter Pfister. Quantart: Quantizing image style transfer towards high visual fidelity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5947–5956, 2023. 5

[9] Jing Huo, Shiyin Jin, Wenbin Li, Jing Wu, Yu-Kun Lai, Yinghuan Shi, and Yang Gao. Manifold alignment for semantically aligned style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14861–14869, 2021. 1

[10] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[11] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10051–10060, 2019. 1

[12] Bo Li, Caiming Xiong, Tianfu Wu, Yu Zhou, Lun Zhang, and Rufeng Chu. Neural abstract style transfer for chinese traditional painting. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part II 14*, pages 212–227. Springer, 2019. 5

[13] Jing Liao, Yuan Yao, Lu Yuan, Gang Hua, and Sing Bing Kang. Visual attribute transfer through deep image analogy. *arXiv preprint arXiv:1705.01088*, 2017. 1

[14] Ming Lu, Hao Zhao, Anbang Yao, Yurong Chen, Feng Xu, and Li Zhang. A closed-form solution to universal style transfer. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5951–5960. IEEE. 5

[15] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5880–5888, 2019. 1

[16] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011. 5

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[18] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[20] Gemma Canet Tarrés, Dan Ruta, Tu Bui, and John Collomosse. Parasol: Parametric style control for diffusion image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2432–2442, 2024. 5

[21] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2

[22] Zhizhong Wang, Lei Zhao, Sihuan Lin, Qihang Mo, Huiming Zhang, Wei Xing, and Dongming Lu. Glstylenet: exquisite style transfer combining global and local pyramid features. *IET Computer Vision*, 14(8):575–586, 2020. 1

[23] Zhizhong Wang, Zhanjie Zhang, Lei Zhao, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Aesust: towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1095–1106, 2022. 5

[24] Zhizhong Wang, Lei Zhao, Haibo Chen, Ailin Li, Zhiwen Zuo, Wei Xing, and Dongming Lu. Texture reformer: Towards fast and universal interactive texture transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2624–2632, 2022. 1

[25] Linfeng Wen, Chengying Gao, and Changqing Zou. Cap-vstnet: Content affinity preserved versatile style transfer. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18300–18309, 2023. 5

[26] Zijie Wu, Zhen Zhu, Junping Du, and Xiang Bai. Ccpl: Contrastive coherence preserving loss for versatile style transfer. In *European Conference on Computer Vision*, pages 189–206. Springer, 2022.

[27] Alice Xue. End-to-end chinese landscape painting creation using generative adversarial networks. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 3863–3871, 2021.

[28] Yuxin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. A

unified arbitrary style transfer framework via adaptive contrastive learning. *ACM Trans. Graph.*, 42(5), 2023.

[29] Zhanjie Zhang, Quanwei Zhang, Wei Xing, Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Junsheng Luan, Yiling Huang, and Huaizhong Lin. Artbank: Artistic style transfer with pre-trained diffusion model and implicit style prompt bank. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7396–7404, 2024. 5

[30] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 5