

AesPA-Net: Aesthetic Pattern-Aware Style Transfer Networks

Kibeom Hong^{1,3,4} Seogkyu Jeon¹ Junsoo Lee⁴ Namhyuk Ahn⁴ Kunhee Kim^{2,4}

Pilhyeon Lee¹ Daesik Kim⁴ Youngjung Uh¹ Hyeran Byun^{1*}

¹Yonsei University ²KAIST AI ³SwatchOn ⁴NAVER WEBTOON AI

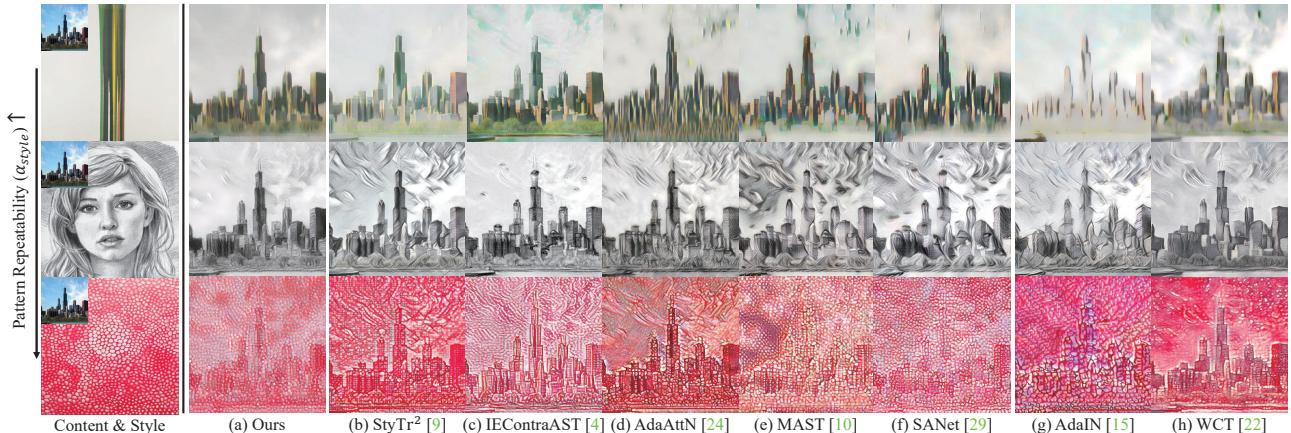


Figure 1. Stylization results. Given input pairs (a fixed content and three references), we compare our method (a) with both attention-based transfer methods (b-f) and global statistic-based methods (g-h). While previous works fail to generate plausible results with various style, our model successfully conducts artistic style transfer well regardless of different style pattern. (Best viewed in zoomed and color.)

Abstract

To deliver the artistic expression of the target style, recent studies exploit the attention mechanism owing to its ability to map the local patches of the style image to the corresponding patches of the content image. However, because of the low semantic correspondence between arbitrary content and artworks, the attention module repeatedly abuses specific local patches from the style image, resulting in disharmonious and evident repetitive artifacts. To overcome this limitation and accomplish impeccable artistic style transfer, we focus on enhancing the attention mechanism and capturing the rhythm of patterns that organize the style. In this paper, we introduce a novel metric, namely pattern repeatability, that quantifies the repetition of patterns in the style image. Based on the pattern repeatability, we propose **Aesthetic Pattern-Aware** style transfer Networks (AesPA-Net) that discover the sweet spot of local and global style expressions. In addition, we propose a novel self-supervisory task to encourage the attention mechanism

to learn precise and meaningful semantic correspondence. Lastly, we introduce the patch-wise style loss to transfer the elaborate rhythm of local patterns. Through qualitative and quantitative evaluations, we verify the reliability of the proposed pattern repeatability that aligns with human perception, and demonstrate the superiority of the proposed framework. All codes and pre-trained weights are available at Kibeom-Hong/AesPA-Net.

1. Introduction

Style transfer aims to render the content of a source image using the elements of a style image. Although the concepts of content and style cannot be rigorously defined, the research community has been building agreeable ideas, considering repeating elements such as brush strokes, color maps, patterns, and certain dominant shapes to be style [11]. Since Gatys et al. [11] pioneered the Neural Style Transfer by using the Gram matrix of deep feature activations as the style representation, previous works have shown surprising performance in Artistic Style Transfer (AST) by employing the global transformation methods [22, 15, 2] and patch-wise swapping approaches [5, 33, 17] that leverage

* Corresponding author
✉ First author email

the similarity between patches in the content and style inputs. However, a major challenge for these methods is the lack of semantic information on the content images, which often leads to distortion artifacts (Fig. 1 (g-h)).

To handle this, recent advances in AST methods [28, 10, 24, 25, 9, 40] have incorporated the attention mechanism which aggregates elements from a style image to render the content of a source image according to semantic correspondence between local patches of the images. While the principle of attention networks has shown the great potential in guiding detailed expression and maintaining content information, they often produce abnormal patterns or evident artifacts, *e.g.*, smudges or floating eyes (Fig. 1 (d-f)). Although several works [9, 4, 37] have attempted to alleviate this problem by employing transformer or external learning strategies, achieving detailed textures such as pointillism or brush strokes remains challenging (Fig. 1 (b-c)).

In order to overcome above issues, we conjecture pitfalls of the attention mechanism for AST. Firstly, we point out that the low semantic correspondence between arbitrary content and style images induces attention mechanisms to focus on limited regions of the style image. This can hinder attention-based methods from accurately capturing and expressing the entire style of the reference images. Consequently, this obstacle leads to disharmonious artifacts since they heavily rely on few style patches such as eyes instead of a representative style *e.g.*, a pencil sketch in a portrait (2nd row of Fig. 1). Furthermore, attention mechanisms utilize only small patches from the style image where large patches would be more suitable. This leads to overly repetitive patterns, even with less-repetitive styles. For instance, despite a simple stripe style, global regions such as the sky are prone to contain artifacts and objects can be painted with excessively repetitive stripes (1st row of Fig. 1).

To this end, we introduce remedies for more delicate artistic expression and improving the attention mechanism. First, we revisit the definition of style considering its unique repeatability. As shown in Fig. 2, our motivation is based on the observation that every style can be expressed as a repetition of appropriate patches which could describe the entire style. To this end, we propose a novel metric *pattern repeatability* which quantifies the frequency of repeating patches in a style image. Then it indicates whether we should bring more effects from attention-based stylization whose advantage lies in details, or global statistic-based stylization whose advantage lies in the smooth reflection of entire styles. Accordingly, we introduce the **Aesthetic Pattern-Aware** style transfer Networks (*AesPA-Net*) which calibrate the stylized features from the two approaches upon the pattern repeatability.

In addition, we propose the self-supervisory task for encouraging the attention module to capture the broader corresponding regions of style images even for arbitrary pairs.

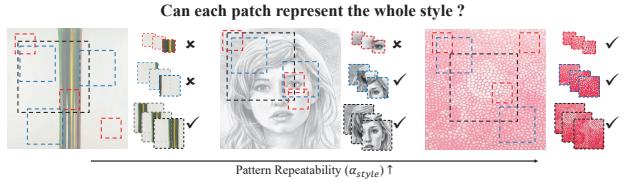


Figure 2. Illustration of our motivation. We show whether each local pattern with various scale could represent the entire style images or not. Motivated by this, we introduce a new metric to quantify the repetition of local patches, *i.e.*, pattern repeatability.

This auxiliary task amounts to maximizing the similarity between the augmented input and its original at the feature level. As a result, it effectively reinforces the attention modules to conduct the artistic stylization. Last but not least, we modify the style loss to be computed between patches with the proper size. Our patch-wise style loss induces the stylized results to reflect the rhythm of local patterns according to the proposed pattern repeatability.

In experiments, we qualitatively show that our networks are capable of articulate artistic stylization with various patterns of style. Besides, through quantitative proxy metrics and the user study, we demonstrate that our framework, *AesPA-Net*, outperforms the state-of-the-art AST studies: five attention-based methods [28, 10, 24, 4, 9] and two global statistics-based methods [22, 15]. Furthermore, we verify that the reliability of pattern repeatability closely aligns with human perception, and provide ablation studies to validate the effect of individual components.

2. Related works

2.1. Artistic Style Transfer

Global transformation and patch-wise swapping-based. Since Gatys *et al.* [11] opened up the research of Neural Style Transfer (NST), NST studies have been explored in various domains such as photo-realistic [42, 39, 3], video [14, 36, 6, 8], and 3D [12, 27]. Among others, Artistic Style Transfer (AST) [22, 15, 5, 33] has drawn huge and broad attention by sublimating daily photos into artistic expressions. Recently, Li *et al.* [22] and Huang *et al.* [15] introduce the Whitening and Coloring Transformation (WCT) and the Adaptive Instance Normalization (AdaIN) to achieve arbitrary style transfer by washing away the global feature statistics and replacing them with the target ones. Meanwhile, Chen *et al.* [5] swap the patches of content features and Sheng *et al.* [33] propose the style decorator to normalize feature with the most correlated style features. Jing *et al.* [17] have improved semantic matching by employing techniques such as Graph Neural Networks. However, these AST methods have difficulty in preserving the original context due to the lack of cues from content images, resulting in unrealistic stylization as well as content distortion. To over-

come this limitation, attention-based local patch stylization methods [41, 28, 10, 24, 4, 40, 37, 9] are proposed.

Attention mechanism-based. AAMS [41] and SANet [28] first employ the self-attention mechanism to derive the semantic mapping for artistic style transfer. MAST [10] adopts the attention mechanism to not only stylized features but also to content and style features itself. In addition, AdaAttN [24] proposes the fusion strategy that exploits both the attention mechanism and the global feature statistic-based method. Besides, StyTr² [9] introduces the Transformer and the content-aware positional encoding for delicate style expression. Nonetheless, they suffer from peculiar repetitive artifacts when they confront a style image with irregular patterns, producing abnormal stylization results. IEcontraAST [4], CAST [40], and AesUST [37] have tackled this problem by exploiting the contrastive and adversarial learning but still fail to express the detailed texture. In this paper, we delve into the relation between the stylization result and the degree of pattern repetition and utilize this relation to better translate the artistic rhythm of the style image, achieving elaborate and aesthetic style transfer.

2.2. Relation between Pattern and Style

A pattern appears in various scales in a single image. In other words, a pattern could be described as an entire image or as a very small local patch. Based on this notion, there are a few studies [18, 19, 16, 35, 29] investigating the correlation between pattern and style. Julez *et al.* [18, 19] introduce the texton theory that the human vision system is related to the frequency of repeated patterns. Jing *et al.* [16] and Virtusio *et al.* [35] propose controllable artistic style transfer networks by manipulating the resolution of reference images. They demonstrate that features extracted from various resolutions of the target image can be interpreted as having different styles, which in turn change the stylization result. In addition, SwappingAE [29] proposes the patch discriminator to criticize the texture based on multi-scale patterns and conduct the texture translation. Furthermore, Cheng *et al.* [7] statistically analyze the degree of stylization and introduce the normalized style loss to express proper repeatability. Our motivation is also in the continuation of these studies: “*The unique pattern repeatability in each image defines its own style*”. In this paper, we aim to capture not only the elements of the style (*e.g.*, color or texture) but also their unique repeatability in organizing the artistic rhythm and expression.

3. Proposed Method

In this section, we describe the proposed ***pattern repeatability*** and ***Aesthetic Pattern-Aware style transfer networks*** (AesPA-Net) in detail. Notably, our goal is to render a stylized image (I^{cs}) from the content image (I^c) and the style image (I^s) by transferring the atmosphere of repeata-

bility as well as elements of style. The overview of our method is illustrated in Fig. 3.

3.1. Pattern Repeatability

Motivation. Despite the advances of attention-based AST methods, there still exists a fundamental limitation of attention networks in artistic style transfer. Due to the low semantic correspondence between the content and the artistic style images, the attention module tends to map the content to only a few local patches. Hence, a specific local patch pattern of the style image is repeatedly expressed in the stylized results. These stylization results would be visually satisfactory when the small local patches have high similarity to each other as well as the global style, *e.g.*, pointillism artworks. Otherwise, the results become unnatural and implausible especially when the local patches cannot represent the overall style, *e.g.*, portraits. Intrinsically, the degree of local pattern repetition and the minimal patch size, which can represent the whole image, vary across different styles as shown in Fig. 2. To this end, we propose a novel metric, ***pattern repeatability***, to quantify the rhythm and repetition of local patterns in the style image.

Method. In order to measure the degree of pattern repeatability in the style image I^s , we consider two perspectives: 1) Similarity between each patch and the whole image, *i.e.*, intra-image patch repeatability, and 2) Similarity between patches, *i.e.*, inter-patch repeatability.

Given an image I , we gather features $f_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ from the l^{th} convolutional layer of encoder ϕ , and then obtain N feature patches $\{P_{l,r}^i\}_{i=1}^N$ by spatially dividing each features with ratio r , *i.e.*, $P_{l,r}^i \in \mathbb{R}^{C_l \times \frac{H_l}{r} \times \frac{W_l}{r}}$. To estimate how much local styles resemble the global style, *i.e.*, intra-image patch repeatability α_{intra} , we measure the cosine-similarity between the Gram matrix of each feature patch $P_{l,r}^i$ and the entire feature f_l as follows.

$$\alpha_{intra} = \frac{1}{L} \sum_{l=1}^L \frac{1}{N} \sum_{i=1}^N \frac{\mathcal{G}(P_{l,r}^i) \cdot \mathcal{G}(f_l)}{\|\mathcal{G}(P_{l,r}^i)\| \|\mathcal{G}(f_l)\|}, \quad (1)$$

where $\mathcal{G}(\cdot)$ depicts the Gram matrix calculation and L indicates the number of convolutional layers in the encoder ϕ . A high α_{intra} indicates that the global style appears consistently in its local patches, whereas a low α_{intra} means local styles are distinguishable from the global style.

In addition, we calculate the similarity between textures of different patches, *i.e.*, inter-patch repeatability α_{inter} , to estimate whether particular local patterns appear frequently. This process is formulated as follows.

$$\alpha_{inter} = \frac{1}{L} \sum_{l=1}^L \frac{1}{|\mathcal{B}|} \sum_{\forall (i,j) \in \mathcal{B}} \frac{\mathcal{G}(P_{l,r}^i) \cdot \mathcal{G}(P_{l,r}^j)}{\|\mathcal{G}(P_{l,r}^i)\| \|\mathcal{G}(P_{l,r}^j)\|}. \quad (2)$$

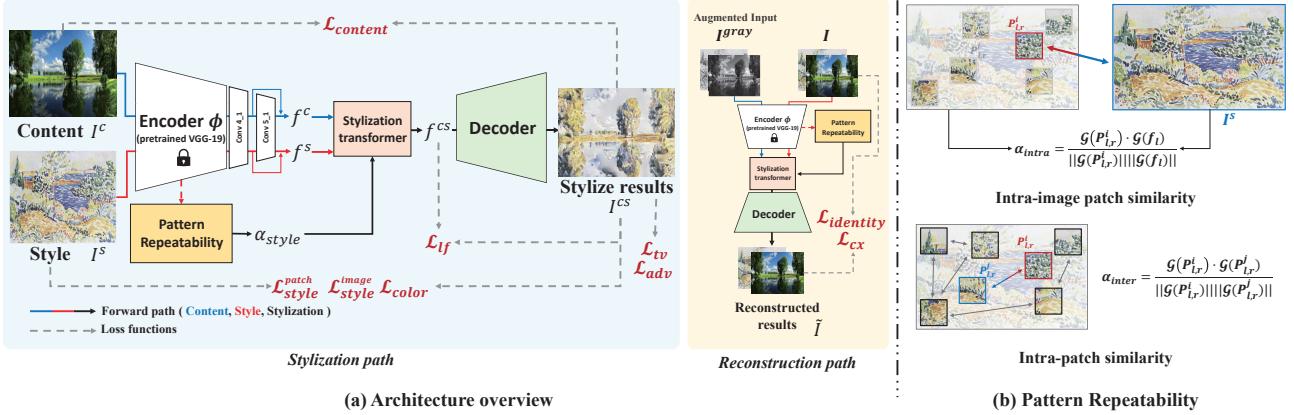


Figure 3. The overview of the aesthetic pattern-aware style transfer networks (AesPA-Net). (a) shows the architecture of AesPA-Net and training procedures, *i.e.*, stylization task (blue box) and reconstruction task (yellow box). (b) depicts the detail of pattern repeatability.

Here \mathcal{B} is the selected pair set out of $N C_2$ pairs with a probability of p for efficient calculation, *i.e.*, $|\mathcal{B}| = p \cdot N C_2$. Moreover, we compute the pattern repeatability with the grayscaled style image in a similar way to capture the pattern repeatability of the regional structure. Finally, the pattern repeatability is computed as:

$$\alpha_{style} = \frac{1}{4} \cdot (\alpha_{inter}^{RGB} + \alpha_{intra}^{RGB} + \alpha_{inter}^{gray} + \alpha_{intra}^{gray}). \quad (3)$$

We depict the conceptual illustration of pattern repeatability in Fig. 3 (b). Detailed analysis on the reliability of pattern repeatability is provided in Sec. 4.1.1.

3.2. Aesthetic Pattern-Aware Style Transfer

We design a novel style transfer networks (AesPA-Net) that can fully reflect the target style by taking its pattern repeatability into account. In order to achieve impeccable artistic stylization with the styles of arbitrary pattern rhythm, the stylization transformer is the key component that adaptively transfers local and global textures in accordance with the pattern repeatability α_{style} . Intuitively, the ideal solution for the style with low pattern repeatability would be directly increasing the input patch size of attentional style transfer. Nevertheless, this is impractical because the memory requisition increases quadratic according to the patch scale. As a result, attention maps for stylization are computed at limited levels *i.e.*, 4th and 5th levels of the encoder. To overcome this limitation, we implicitly employ a statistics-based transformation as a practical complement to global stylization. Concretely, the global stylized feature f_{global}^{cs} can capture unattended local patterns by matching global statistics to style features of I^S . Meanwhile, the features f_{attn}^{cs} from attentional style transfer module emphasize important local patterns and compensates for the distorted content that is caused during whitening.

Consequently, our stylization transformer is composed of an attentional style transfer module and a global feature

statistic transformation module. From the patch-wise attentional style transfer module, we obtain the stylized feature f_{attn}^{cs} based on local patch-wise correspondence [24] as:

$$\begin{aligned} f_{attn}^{cs} &= AdaAttN(f_l^c, f_l^s, f_{1:l}^c, f_{1:l}^s), \\ f_{attn}^{cs} &= h(f_4^{cs} + \psi_{\uparrow}(f_5^{cs})), \end{aligned} \quad (4)$$

where l denotes the index of layer, $h(\cdot)$ and ψ_{\uparrow} indicate the convolutional layer and upsampling function, respectively.

To acquire f_{global}^{cs} , we apply the global statistic transformation [22] on conv4_1 features. After that, we obtain the final stylized feature f^{cs} as:

$$f^{cs} = \alpha_{style} \cdot f_{attn}^{cs} + (1 - \alpha_{style}) \cdot f_{global}^{cs}, \quad (5)$$

where f_{attn}^{cs} and f_{global}^{cs} represent the attended stylized features and global statistic transformed features, respectively. α_{style} is the overall pattern repeatability obtained in Eq. 3. Finally, our decoder inverts the stylized feature maps to an artistic result I^{cs} . In the supplementary material, we show that our stylization transformer also works effectively with another global transformation *i.e.*, AdaIN [15].

As illustrated in Fig. 3 (a), AesPA-Net consists of an encoder, a stylization transformer, and a decoder. We employ the pre-trained VGG-19 [34] as an encoder, while the decoder is designed to mirror the architecture of the encoder by replacing down-sampling operations with up-sampling ones.

3.3. Patch-wise Style Loss

To further convey the aesthetic expression of style into the content, we design the patch-wise style loss ($\mathcal{L}_{style}^{patch}$). Compared to the conventional style loss ($\mathcal{L}_{style}^{image}$) that targets to bring the global style, the proposed patch-wise style loss aims to transfer the elaborate rhythm of local patterns. As shown in Fig. 4, we patchify the images into

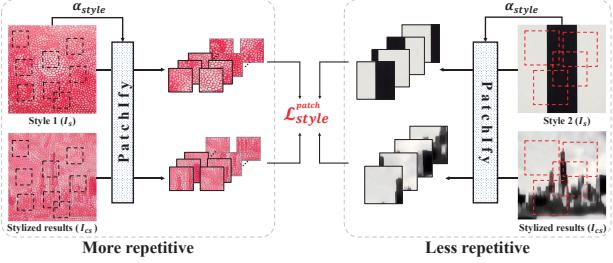


Figure 4. The illustration of patch-wise style loss $\mathcal{L}_{style}^{patch}$.

M patches $\{\bar{I}_m\}_{m=1}^M$ by spatially dividing them with the scale s , *i.e.*, $\bar{I}_m \in \mathbb{R}^{3 \times \frac{H}{s} \times \frac{W}{s}}$. We define the scale s as $s = \max(2^{8 \cdot \alpha_{style} - 5}, 1)$ to reflect the pattern repeatability. Afterward, we calculate the L_2 distance between the Gram matrices of patches from style and stylized results as follows.

$$\mathcal{L}_{style}^{patch} = \frac{1}{M} \sum_{m=1}^M \sum_{l=4}^5 \|\mathcal{G}(\phi_l(\bar{I}_m^{cs})) - \mathcal{G}(\phi_l(\bar{I}_m^s))\|_2, \quad (6)$$

where $\phi_l(\cdot)$ denotes the activation maps after ReLU_{l-1} layers of encoder. Note that we utilize the centered-Gram matrix [35] which could capture the channel-wise negative correlation. We verify the effectiveness of the proposed loss function in Sec. 4.4.

3.4. Improving Attention Module for AST

We investigate the training procedure of attention mechanisms employed in previous AST methods [28, 10]. They train the attention networks to solve the reconstruction task by satisfying the identity loss. However, we insist that this loss design is inappropriate for AST since the content, style, and reconstruction target images are totally identical. As the query features from the content already contain sufficiently rich information for reconstruction, key features from the style are likely to be ignored, thus hindering the attention module from learning useful semantic correspondence between the content and the style.

To alleviate this problem, we introduce an effective self-supervisory task for correspondence learning. We input the original image I and its grayscaled version I^{gray} as the style (key) and the content (query) images, respectively. In this way, due to the reduced amount of information comes from the augmented query, the attention module is required to precisely capture the contextual relationship from reference images. Finally, the loss functions of the self-supervisory reconstruction path are formulated as follows.

$$\begin{aligned} \mathcal{L}_{rec} &= \lambda_{identity} \mathcal{L}_{identity} + \lambda_{cx} \mathcal{L}_{cx}, \text{ where} \\ \mathcal{L}_{identity} &= \sum_{l=2}^5 \|\phi_l(\tilde{I}) - \phi_l(I)\|_2, \\ \mathcal{L}_{cx} &= CX(\tilde{I}, I), \end{aligned} \quad (7)$$

where \tilde{I} is the reconstructed image with I and I^{gray} . $\phi_l(\cdot)$ describes the activation maps after ReLU_{l-1} layers of the encoder, *i.e.*, VGG-19. \mathcal{L}_{cx} depicts the contextual similarity loss [26].

Note that, we modify the reconstruction loss \mathcal{L}_{rec} to be computed with deep features for relieving blurriness and amplifying perceptual quality. The overall reconstruction path is depicted in Fig. 3 (a). In Sec. 4.1.2 and 4.4, we show the validity of the proposed strategy via ablations studies.

3.5. Training

As illustrated in Fig. 3, the decoder and the stylization transformer are trained to perform two different tasks (*i.e.*, reconstruction and stylization) in an end-to-end manner. In the reconstruction path, our networks are encouraged to reconstruct the original RGB images when their grayscaled ones are given as the content. Recapitulating the improved attention module in Sec. 3.4, we induce the stylization transformer to learn semantic correspondence between the content and the style images.

In the stylization path, the style loss for aesthetic style representation is derived by the weighted sum of losses as:

$$\begin{aligned} \mathcal{L}_{style} &= \lambda_{image} \mathcal{L}_{style}^{image} + \lambda_{patch} \mathcal{L}_{style}^{patch} + \lambda_{lf} \mathcal{L}_{lf} \\ &\quad + \lambda_{content} \mathcal{L}_{content} + \lambda_{color} \mathcal{L}_{color} + \lambda_{tv} \mathcal{L}_{tv}. \end{aligned} \quad (8)$$

Notably, the style loss $\mathcal{L}_{style}^{image}$ and the proposed patch-wise style loss $\mathcal{L}_{style}^{patch}$ contribute to global and local target texture expression, respectively. Besides, the content loss $\mathcal{L}_{content}$ is adopted for the content as well as context preservation. In addition, we adopt the local feature loss \mathcal{L}_{lf} [24] to encourage feature consistency. Moreover, we utilize the color loss \mathcal{L}_{color} [1] to follow the color palettes of the style image, and the total variation loss \mathcal{L}_{tv} to encourage pixel-wise smoothness. Each loss is described as follows.

$$\begin{aligned} \mathcal{L}_{style}^{image} &= \sum_{l=4}^5 \|\mathcal{G}(\phi_l(I^{cs})) - \mathcal{G}(\phi_l(I^s))\|, \\ \mathcal{L}_{content} &= \sum_{l=4}^5 \|\phi_l(I^{cs}) - \phi_l(I^c)\|, \\ \mathcal{L}_{lf} &= \sum_{l=4}^5 \|\phi_l(I^{cs}) - f_{attn}^{cs}\|, \\ \mathcal{L}_{color} &= \frac{1}{\sqrt{2}} \left\| (H^s)^{1/2} - (H^{cs})^{1/2} \right\|, \end{aligned} \quad (9)$$

where $H^{1/2}$ denotes the element-wise square root of the color histogram. To further elevate the stylization performance, we employ the multi-scale discriminator with the adversarial loss (\mathcal{L}_{adv}) [13].

Finally, the full objective of our model is summation of those loss functions: $\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{style} + \lambda_{adv} \mathcal{L}_{adv}$.

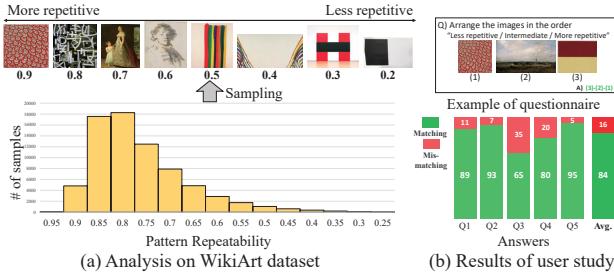


Figure 5. Analysis of the proposed pattern repeatability α_{style} . (a) presents the distribution of WikiArt [31] dataset and examples according to α_{style} . (b) shows the results of user study on correlation of α_{style} with the human perception.

4. Experiments

Implementation details. We train AesPA-Net on MS-COCO [23] and WikiART [31] datasets, each containing about 120K real photos and 80K artistic images. We use the Adam [20] optimizer to train the attention module as well as the decoder with a batch size of 6, and the learning rate is set to 1e-4. During the training and inference, we rescale the input images to have a spatial resolution of 256×256 . We set the weighting factors of loss functions as follows: $\lambda_{identity} = 1$, $\lambda_{cx} = 1$, $\lambda_{content} = 1$, $\lambda_{image} = 10$, $\lambda_{lf} = 100$, $\lambda_{patch} = 0.5$, $\lambda_{color} = 1$, $\lambda_{adv} = 0.1$, $\lambda_{tv} = 1$. We implement our framework with PyTorch [30] and train our model using a single GTX 3090Ti.

4.1. Analysis

4.1.1 Reliability of the pattern repeatability

We analyze the proposed pattern repeatability (α_{style}) in order to validate its reliability. Firstly, we analyze the WikiArt [31] dataset by computing the pattern repeatability of each artwork to examine whether the extracted α_{style} matches well with the visual pattern rhythm of the image. We depict the histogram of pattern repeatability in Fig. 5 (a). The mean of pattern repeatability is about 0.79 and the standard deviation is 0.1. The highest and lowest value of α_{style} is 0.97 and 0.26, respectively.

For qualitative verification of pattern repeatability, we visualize randomly sampled images for every 0.1 range. As shown in the upper part of Fig. 5 (a), images with high α_{style} have a uniformly repeated local pattern, and the local patterns thus can express the global pattern regardless of its scale. On the other hand, images with low α_{style} tend to have irregular local patterns, and their style hence should be interpreted from the entire image.

To further demonstrate that the proposed pattern repeatability aligns well with human perception, we conduct the user study. We compose each questionnaire with three randomly shuffled images of different pattern repeatability α_{style} from ‘Less repetitive’ (0.1~0.3) to ‘More repetitive’ (0.8~0.9). Then, we asked the participants to arrange the images in ascending order to compare the responses with the sequence according to the proposed metric. In Fig. 5 (b), the answers mostly agree with the sequence of the proposed metric, reporting the average matching rate of 84%. This verifies that α_{style} reflects the pattern repetition as the perception of the human vision system.

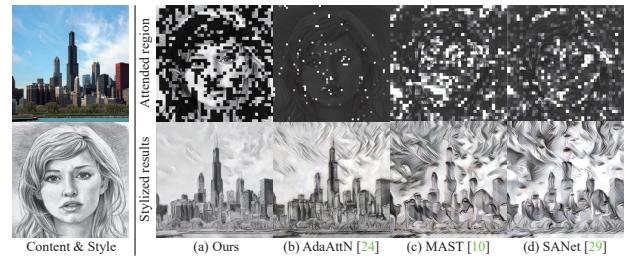


Figure 6. Visualization of attended region (first row) and stylized results (second row) given a pair of content and style images.

tive” (0.8~0.9). Then, we asked the participants to arrange the images in ascending order to compare the responses with the sequence according to the proposed metric. In Fig. 5 (b), the answers mostly agree with the sequence of the proposed metric, reporting the average matching rate of 84%. This verifies that α_{style} reflects the pattern repetition as the perception of the human vision system.

4.1.2 Effectiveness of improved attention module

We demonstrate the effectiveness of the improved attention module compared to the previous attention-based baselines [28, 10, 24]. The proposed self-supervisory task reinforces the overall style transmission effect. As shown in Fig. 6, the attention module from previous works [28, 10, 24] selects some limited local patches (e.g., eyes), leading to disharmonious artifacts in stylized results. On the other hand, our proposed grayscaled augmentation strategy helps the attention module to capture meaningful and broader correspondence between the content and the style image. Finally, AesPA-Net generates plausible artistic outputs without disharmonious artifacts.

4.2. Qualitative Results

As shown in Fig. 7, we qualitatively compare our method with five state-of-the-art attention-based AST models [9, 4, 24, 10, 28] and two global statistic-based methods [15, 22]. Specifically in the 1st and 2nd rows, most works [9, 24, 10, 28, 15] generate repetitive artifacts in stylized results, e.g., wave stains on the sky region. Although these artifacts are rarely appearing in IEContraAST [4], it still produces disharmonious distortions of eyes around the Eiffel tower and struggles to transfer detailed texture and strokes of Van Gogh’s painting. On the other hand, AesPA-Net successfully reproduces the contents respecting the texture and color palette of the artworks without artifacts.

With style images of high pattern repeatability (3rd, 4th, and 5th rows), our proposed methods infuse the delicate rhythm of target patterns into the content images. Particularly, the glass surface is expertly brushed with vivid yellow strokes of the bridge, and the urbanscape is adeptly illustrated with Mondrian’s signature irregular rectangles.

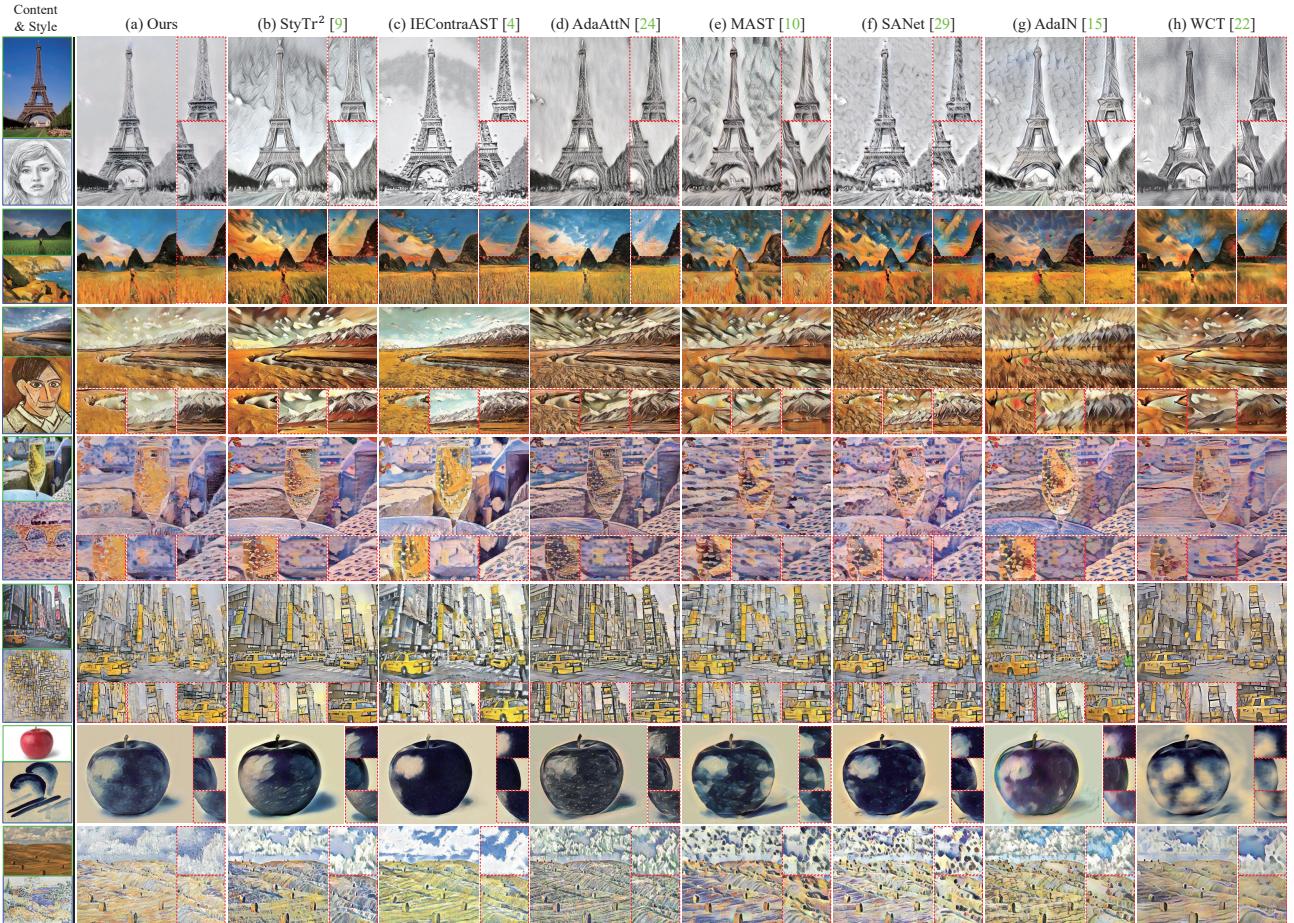


Figure 7. Qualitative comparisons with state-of-the-art AST methods. Green and blue boxes indicate the content and reference images respectively. We provide zoomed patches (dashed red boxes) for better comparison.

	AesPA-Net (Ours)	StyTr ² [9]	IIEContraAST [4]	AdaAttN [24]	MAST [10]	SANet [28]	AdaIN [15]	WCT [22]
User preference	0.65*	0.65 / 0.35	0.70 / 0.30	0.67 / 0.33	0.58 / 0.42	0.68 / 0.32	0.57 / 0.43	0.72 / 0.28
Human deception rate	0.72*	0.63 / 0.37	0.68 / 0.32	0.63 / 0.37	0.77 / 0.23	0.74 / 0.26	0.70 / 0.30	0.87 / 0.13
Content fidelity (\uparrow) [38]	0.690	0.685	0.698	0.645	0.554	0.567	0.609	0.437
Style loss (\downarrow) [11]	0.258	0.305	0.316	0.278	0.274	0.313	0.284	0.345
Pattern difference (\downarrow)	0.075	0.109	0.113	0.100	0.104	0.097	0.091	0.109
Inference time (Sec.)	0.065	0.445	0.381	0.054	0.020	0.015	0.349	1.071

Table 1. Quantitative comparisons with state-of-the-art AST methods. * denotes the average user responses.

Moreover, the hills and sky are represented through skillful stippling. However, some methods produce photo-realistic drawings (b-d) or fall short of representing local patterns of the style image (e-h).

Lastly, with less-repetitive styles (6th and 7th rows), the proposed method effectively describes the global representation without mottled artifacts. Notably, patches of the stylized apple most closely resemble the style of a given single brushstroke drawing. To summarize, AesPA-Net synthesizes artistic stylized results with high fidelity in terms of texture, color, and pattern repeatability across a wide range of style patterns. We provide additional qualitative comparisons in the supplementary material.

4.3. Quantitative Results

User study. Since judging successful artistic stylization is quite subjective, we conducted a user study to compare the quality of artistic expression (*i.e.*, user preference) and the reality of stylized results (*i.e.*, human deception rate). We sampled 42 stylized results from random pairs of content and style images, and conduct one-to-one comparisons with representative AST models [28, 10, 24, 4, 9, 22, 15]. We collected 2310 responses from 55 subjects.

As reported in the 1st and 2nd rows of Tab. 1, most participants favored our AesPA-Net over competitors in terms of artistic representation. Plus, when participants are asked

to choose more realistic artwork following previous works [32, 21], AesPA-Net surpassed previous methods with the evident gap in terms of the human deception rate.

Statistics. Following previous works [37, 38, 13], we respectively compute degrees of artistic stylization and contextual preservation of content information via two proxy metrics: style loss [11] and content fidelity (CF) [38, 37]. To calculate CF, we estimate the cosine-similarity between features of content and stylized images as follows: $CF(I^{cs}, I^c) = \frac{1}{L} \sum_{l=1}^L \frac{f_l^{cs} \cdot f_l^c}{\|f_l^{cs}\| \|f_l^c\|}$, where f_l^c and f_l^{cs} respectively are content and stylized features extracted from l^{th} convolutional layer of an encoder. Additionally, we compute the style loss [11] between artistic stylized outputs and reference samples over five crop patches, following DSTN [13]. We sample 65 random pairs of content and style references for quantitative evaluation.

The 3rd and 4th rows of Tab. 1 depict the quantitative results compared with previous AST methods. Noticeably, they fall behind either in preserving the original context (e.g., WCT [22], SANet [28], MAST [10]), or in expressing artistic styles (e.g., IEcontraAST [4], StyTr² [9]). In contrast, AesPA-Net achieves state-of-the-art results in terms of the style loss while being competent in the content fidelity. To further investigate the scalability of our proposed metric, *pattern repeatability*, we assess the quality of AST by quantifying the $L1$ -distance of pattern repeatability between reference and stylized images as follows: $\|\alpha_{style} - \alpha_{stylization}\|_1$. As shown in the 5th arrow of Tab. 1, AesPA-Net transfers the specific pattern well and the pattern repeatability could be a new metric for AST.

Inference time. We record the inference time with a input resolution of 512×512 compared with other methods in the last row of Tab. 1. The inference speed of our AesPA-Net is comparable to state-of-the-art methods, and therefore it can be adopted in practice for real-time image stylization.

4.4. Ablation Studies

In Fig. 8, we demonstrate the effectiveness of each proposed component of AesPA-Net qualitatively: the improved attention module training (b), stylization transformer (c), and the patch-wise style loss (d). Compared to the attention-based baseline [24], our proposed self-supervisory reconstruction task helps the attention module to capture meaningful and broader correspondence regions from the style image. For example in the first row, the baseline fills the texture of a water tap with the stripe pattern, whereas the improved baseline paints the water tap with black as the style image. Also in the last row, artifacts of repetitive eye patterns are removed in (b). As the proposed grayscale augmentation strategy guides the attention module to find dense semantic correspondence, the local textures in the style image are delivered to the proper regions in the content.

Besides, the proposed stylization transformer removes

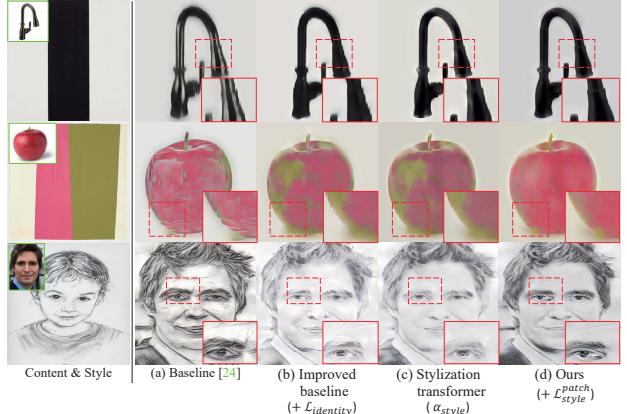


Figure 8. Qualitative ablation studies of AesPA-Net. From the baseline (a) to the full model (d), we show the effectiveness of proposed components sequentially. The green box indicates the input content image and red box depicts the zoomed patches.

defective artifacts caused by unnatural mapping from the attention module. Particularly, the aliased outline of a water tap is smoothed and the noisy patterns in the apple are cleared out. As the stylization transformer adaptively fuses the stylized features f_{attn}^{cs} and f_{global}^{cs} in accordance with the α_{style} , it relieves the artifacts by reflecting the global style pattern. Also, the features from the attention module f_{attn}^{cs} prevent the exaggerated distortion of the original context.

Lastly, the patch-wise style loss boosts the overall style expression and reduces hallucinations in the skin, as shown in Fig. 8 (d). Since we utilize α_{style} to find the minimal patch \bar{I}_m^s that can express the global style image by its repetition, its feature gram matrix $\mathcal{G}(\phi_l(\bar{I}_m^s))$ represents a unit style of the global pattern rhythm. Therefore, exploiting it as a guideline results in delicate and aesthetic stylization.

5. Conclusion

In this paper, we revisit the definition of style with pattern repeatability. The proposed pattern repeatability quantifies the rhythm and repetition of local patterns in the style images by taking into account inter-patch similarity and intra-image patch similarity. Based on the pattern repeatability, we designed a novel style transfer framework, *AesPA-Net*, that can deliver the full extent of the target style via the stylization transformer. Extensive experiments show that our AesPA-Net successfully conducts elaborate artistic style transfer with various style patterns. Our future goal is to expand our method for more impeccable artistic style transfer by considering not only repeatability but also the context in the style image, such as relative locations of style patches.

Acknowledgements. This project was supported by the National Research Foundation of Korea grant funded by the Korean government (MSIT) (No. 2022R1A2B5B02001467).

References

[1] Mahmoud Afifi, Marcus A. Brubaker, and M. S. Brown. Histogan: Controlling colors of gan-generated and real images via color histograms. *CVPR*, pages 7937–7946, 2021. 5

[2] Jie An, Siyu Huang, Yibing Song, Dejing Dou, Wei Liu, and Jiebo Luo. Artflow: Unbiased image style transfer via reversible neural flows. In *CVPR*, pages 862–871, 2021. 1

[3] Jie An, Haoyi Xiong, Jun Huan, and Jiebo Luo. Ultrafast photorealistic style transfer via neural architecture search. In *AAAI*, 2020. 2

[4] Haibo Chen, Lei Zhao, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, and Wei Xing. Artistic style transfer with internal-external learning and contrastive learning. In *NeurIPS*, 2021. 2, 3, 6, 7, 8

[5] Tian Qi Chen and Mark Schmidt. Fast patch-based style transfer of arbitrary style. In *Workshop in Constructive Machine Learning, NIPS*, 2016. 1, 2

[6] Xinghao Chen, Yiman Zhang, Yunhe Wang, Han Shu, Chun-jing Xu, and Chang Xu. Optical flow distillation: Towards efficient and stable video style transfer. In *ECCV*, pages 614–630, 2020. 2

[7] Jiaxin Cheng, Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and P. Natarajan. Style-aware normalized loss for improving arbitrary style transfer. *CVPR*, pages 134–143, 2021. 3

[8] Yingying Deng, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, and Changsheng Xu. Arbitrary video style transfer via multi-channel correlation. In *AAAI*, 2021. 2

[9] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *CVPR*, pages 11326–11336, 2022. 2, 3, 6, 7, 8

[10] Yingying Deng, Fan Tang, Weiming Dong, Wen-Cheng Sun, Feiyue Huang, and Changsheng Xu. Arbitrary style transfer via multi-adaptation network. In *ACM Multimedia*, pages 2719–2727, 2020. 2, 3, 5, 6, 7, 8

[11] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 1, 2, 7, 8

[12] Lukas Höllerin, Justin Johnson, and Matthias Nießner. Stylemesh: Style transfer for indoor 3d scene reconstructions. *CVPR*, pages 6188–6198, 2022. 2

[13] Kibeam Hong, Seogkyu Jeon, Huan Yang, Jianlong Fu, and Hyeran Byun. Domain-aware universal style transfer. *ICCV*, pages 14589–14597, 2021. 5, 8

[14] Haozhi Huang, Hao Wang, Wenhan Luo, Lin Ma, Wenhao Jiang, Xiaolong Zhu, Zhifeng Li, and W. Liu. Real-time neural style transfer for videos. *CVPR*, pages 7044–7052, 2017. 2

[15] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017. 1, 2, 4, 6, 7

[16] Yongcheng Jing, Yang Liu, Yezhou Yang, Zunlei Feng, Yizhou Yu, Dacheng Tao, and Mingli Song. Stroke controllable fast style transfer with adaptive receptive fields. In *ECCV*, pages 238–254, 2018. 3

[17] Yongcheng Jing, Yining Mao, Yiding Yang, Yibing Zhan, Mingli Song, Xinchao Wang, and Dacheng Tao. Learning graph neural networks for image style transfer. In *ECCV*, pages 111–128, 2022. 1, 2

[18] B. Julesz. Visual pattern discrimination. *IRE Transactions on Information Theory*, 8(2):84–92, 1962. 3

[19] Béla Julesz. A brief outline of the texton theory of human vision. *Trends in Neurosciences*, 7:41–45, 1984. 3

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[21] Dmytro Kotovenko, Artsiom Sanakoyeu, Pingchuan Ma, Sabine Lang, and Björn Ommer. A content transformation block for image style transfer. *CVPR*, pages 10024–10033, 2019. 8

[22] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NeurIPS*, pages 386–396, 2017. 1, 2, 4, 6, 7, 8

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 6

[24] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *ICCV*, pages 6629–6638, 2021. 2, 3, 4, 5, 6, 7, 8

[25] Xuan Luo, Zhen Han, Lingkang Yang, and Lingling Zhang. Consistent style transfer. *ArXiv*, abs/2201.02233, 2022. 2

[26] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, pages 768–783, 2018. 5

[27] Fangzhou Mu, Jian Wang, Yichen Wu, and Yin Li. 3d photo stylization: Learning to generate stylized novel views from a single image. *CVPR*, pages 16252–16261, 2022. 2

[28] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *CVPR*, pages 5880–5888, 2019. 2, 3, 5, 6, 7, 8

[29] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *NeurIPS*, pages 7198–7211, 2020. 3

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017. 6

[31] Fred Phillips and Brandy Mackintosh. Wiki art gallery, inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3):593–608, 2011. 6

[32] Artsiom Sanakoyeu, Dmytro Kotovenko, Sabine Lang, and Björn Ommer. A style-aware content loss for real-time hd style transfer. In *ECCV*, 2018. 8

[33] Lu Sheng, Ziyi Lin, Jing Shao, and Xiaogang Wang. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *CVPR*, pages 8242–8250, 2018. 1, 2

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

- [35] John Jethro Virtusio, Daniel Stanley Tan, Wen-Huang Cheng, M. Tanveer, and Kai-Lung Hua. Enabling artistic control over pattern density and stroke strength. *IEEE Transactions on Multimedia*, 23:2273–2285, 2021. [3](#), [5](#)
- [36] Wenjing Wang, Jizheng Xu, Li Zhang, Yue Wang, and Jiaying Liu. Consistent video style transfer via compound regularization. In *AAAI*, 2020. [2](#)
- [37] Zhizhong Wang, Zhanjie Zhang, Lei Zhao, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Aesust: Towards aesthetic-enhanced universal style transfer. In *ACM Multimedia*, 2022. [2](#), [3](#), [8](#)
- [38] Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. Evaluate and improve the quality of neural style transfer. *Comput. Vis. Image Underst.*, 207:103203, 2021. [7](#), [8](#)
- [39] Xide Xia, Meng Zhang, Tianfan Xue, Zheng Sun, Hui Fang, Brian Kulis, and Jiawen Chen. Joint bilateral learning for real-time universal photorealistic style transfer. In *ECCV*, pages 327–342, 2020. [2](#)
- [40] Yu xin Zhang, Fan Tang, Weiming Dong, Haibin Huang, Chongyang Ma, Tong-Yee Lee, and Changsheng Xu. Domain enhanced arbitrary image style transfer via contrastive learning. In *ACM SIGGRAPH*, pages 1–8, 2022. [2](#), [3](#)
- [41] Yuan Yao, Jianqiang Ren, Xuansong Xie, Weidong Liu, Yong-Jin Liu, and Jun Wang. Attention-aware multi-stroke style transfer. In *CVPR*, pages 1467–1475, 2019. [3](#)
- [42] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, pages 9036–9045, 2019. [2](#)