

# AesFA: An Aesthetic Feature-Aware Arbitrary Neural Style Transfer

Joonwoo Kwon<sup>1\*</sup>, Sooyoung Kim<sup>1\*</sup>, Yuewei Lin<sup>2 †</sup>, Shinjae Yoo<sup>2 †</sup>, Jiook Cha<sup>1 †</sup>

<sup>1</sup>Seoul National University

<sup>2</sup>Brookhaven National Laboratory

{pioneers, rlatndud0513, connectome}@snu.ac.kr, {ywlin, sjyoo}@bnl.gov

## Abstract

Neural style transfer (NST) has evolved significantly in recent years. Yet, despite its rapid progress and advancement, existing NST methods either struggle to transfer aesthetic information from a style effectively or suffer from high computational costs and inefficiencies in feature disentanglement due to using pre-trained models. This work proposes a lightweight but effective model, **AesFA**—**Aesthetic Feature-Aware NST**. The primary idea is to decompose the image via its frequencies to better disentangle aesthetic styles from the reference image while training the entire model in an end-to-end manner to exclude pre-trained models at inference completely. To improve the network’s ability to extract more distinct representations and further enhance the stylization quality, this work introduces a new aesthetic feature: contrastive loss. Extensive experiments and ablations show the approach not only outperforms recent NST methods in terms of stylization quality, but it also achieves faster inference. Codes are available at <https://github.com/Sooyoungg/AesFA>.

## Introduction

Neural Style Transfer (NST) is an artistic application that transfers the style of one image to another while preserving the original content. Initially introduced by (Gatys, Ecker, and Bethge 2016), this area has gained substantial momentum with the advancement of deep neural networks. Despite such progress, a significant chasm persists between authentic artwork and synthesized stylizations. Existing NST methods, as shown in Figure 1, struggle to capture essential aesthetic features, such as tones, brushstrokes, textures, grains, and the local structure from style images, leading to discordant colors and irrelevant patterns. Ideally, the goal of using NST is to extract a style from the image and transfer it to content, necessitating representations that capture both image semantics and stylistic changes. This work focuses on defining these *style* representations.

In the context of painting, style representations are defined by attributes, such as overall color and/or the local structure of brushstrokes. Most NST algorithms define style representations as spatially agnostic features to encode this

\*These authors contributed equally.

†Co-corresponding authors.

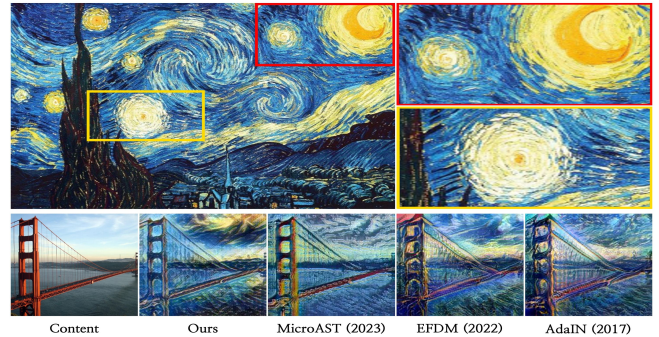


Figure 1: Top: *The Starry Night* by Vincent Van Gogh. The styles have a strong correlation with spatial information, as evidenced by the presence of whirling patterns and expressionistic yellow stars in the “sky”. Bottom: Compared with other NST methods, our method can faithfully transfer styles while ensuring the spatial information.

information. For example, Gatys et al. (Gatys, Ecker, and Bethge 2016) use gram matrices, while Huang et al. (Huang and Belongie 2017) employ mean and variance alignment to obtain a style representation. Despite their success, they rely solely on summary statistics. Thus, they lack spatial information representation. In fact, style representations are *highly correlated* to spatial information. For example, Vincent van Gogh’s *The Starry Night* (Figure 1) has expressionistic yellow stars and a moon that dominate the upper center and right, while dynamic swirls fill the center of the sky. In pondering the *style* of this painting, its focal point primarily resides in the sky rather than the village or cypress trees. Therefore, when transferring *The Starry Night*’s style, the expected style output likely would be the dynamic swirls and expressionistic yellow stars in the sky. From this point of view, spatial information keenly matters in style representations. However, most NST algorithms fail to recognize such distinct spatial styles due to their spatial-independent style representations, leading to stylizations lacking in spatial coherence (refer to the bottom panel in Figure 1).

To enhance stylization, we propose a lightweight yet effective model that we call, **Aesthetic Feature-Aware Arbitrary NST**, or **AesFA**. AesFA overcomes prior NST limitations by encoding style representations while retaining

spatial details. To expedite the extraction of aesthetic features, we decompose the image into two discrete complementary components, i.e., the high- and low-frequency parts. High frequency captures details including textures, grains, and brushstrokes, while low frequency encodes global structures and tones. On the other hand, existing NST algorithms often neglect this disentanglement and extract style features from a mix of irrelevant information. Specifically, we employ Octave Convolution operators (OctConv) (Chen et al. 2019) to decompose and process input images by frequency, which eliminates the need for cumbersome mathematical algorithms like Fast Fourier Transform (FFT) (Gentleman and Sande 1966). This design ensures the model remains lightweight and effective when disentangling features. Furthermore, inspired by adaptive convolutions (AdaConv) (Chandran et al. 2021), which simultaneously blend statistical and structural styles to the contents, we modify AdaConv by effectively combining frequency-decomposed content features with predicted *aesthetic feature-aware kernels and biases*. We refer to the modified stylization module as *Adaptive Octave Convolution (AdaOct)* because it employs AdaConv followed by an OctConv. In AdaOct, frequency-decomposed features undergo convolution with predicted aesthetic-aware kernels and biases, followed by OctConv for exchanging features’ high- and low-frequency components. Consequently, AdaOct achieves superior stylization and reduces unwanted artifacts.

Another challenge is that existing NST methods heavily rely on pre-trained networks, e.g., VGG (Simonyan and Zisserman 2014), for feature extraction. However, using such networks during inference is inefficient because of the computational demands from fully connected layers. This limits NST’s use at high resolutions (e.g., 2K; 4K) and in mobile or real-time scenarios. Larger images also struggle with preserving texture and grains in style transfer. To mitigate this limitation, a prior study (Wang et al. 2023) adopted contrastive learning for end-to-end training while excluding pre-trained convolutional neural networks (CNNs) at inference. However, this approach is computationally expensive and inefficient as it uses all negative samples in a mini-batch, especially with higher-resolution samples. This prompts a question: *are all negative samples necessary?* Intuitively, the more distant negative samples contribute less to training as they are already well discriminated from the positive sample and vice versa. Inspired by hard negative mining, we redefine “negative” samples as the  $k$ -th nearest negative samples to the stylized output, introducing efficient contrastive learning for aesthetic features via pre-trained VGG network.

Overall, AesFA outperforms state-of-the-art algorithms in terms of the structural similarity index (SSIM) and average VGG style perceptual loss across all spatial resolutions, ranging from 256 to 4K. Regardless of image resolution, our method achieves state-of-the-art performance, inferring a single image in under 0.02 seconds.

The contributions of this work are summarized as follows:

- We propose a lightweight yet effective model for aesthetic feature-aware NST, which maintains the spatial style information and decomposes images by frequency to improve feature extraction, substantially enhancing the

stylization quality and computational efficiency at the same time.

- To effectively infuse frequency-decomposed content features with aesthetic features, a new stylization module, AdaOct, is proposed that yields more satisfying stylizations with sophisticated aesthetic characteristics. To further accelerate the networks’ capability to extract more distinct aesthetic representations, a straightforward contrastive learning for aesthetic features also is proposed.
- We show that our method achieves generalization, quality, and efficiency simultaneously across various spatial resolutions by conducting comprehensive comparisons with several state-of-the-art NST methods.

## Related Work

NST emerged with Gatys et al. (Gatys, Ecker, and Bethge 2016), but its optimization is computationally intensive. To deal with this issue, Johnson et al. (Johnson, Alahi, and Fei-Fei 2016) introduced perceptual losses for real-time processing. Subsequent work (Gatys et al. 2017; Ghiasi et al. 2017; Chen and Schmidt 2016; Ulyanov, Vedaldi, and Lempitsky 2017; Dumoulin, Shlens, and Kudlur 2016; Ulyanov et al. 2016) improved NST without sacrificing speed. However, all were limited to specific styles. Huang et al. (Huang and Belongie 2017) proposed Adaptive Instance Normalization (AdaIN) for arbitrary style transfer, which has been extended (Sheng et al. 2018; Kotovenko et al. 2019; Jing et al. 2020; Shen, Yan, and Zeng 2018; Li et al. 2017; Lin et al. 2021; Yoo et al. 2019) for successful style transfer onto content images. Chandran et al. (Chandran et al. 2021) improved AdaIN with AdaConv for structure-aware style transfer. AdaConv simultaneously adapts statistical and structural styles. However, AdaConv’s convolution kernels and biases incur high computational costs. Recent work (An et al. 2023; Wang et al. 2023; Wang, Li, and Vasconcelos 2021) has highlighted drawbacks in NST methods that rely on pre-trained CNNs, e.g., VGG-19 (Simonyan and Zisserman 2014), for feature extraction from the reference image. Wang et al. (Wang, Li, and Vasconcelos 2021) have enhanced non-VGG architectures’ robustness via activation smoothing in stylization loss. An et al. (An et al. 2023) explore alternative architectures, such as GoogLeNet (Szegedy et al. 2015), yet they lack specificity for NST, yielding unsatisfactory stylization outcomes and high memory use. Instead, our objective is productive mobile NST that incorporates aesthetic features.

**Multiscale representation learning.** Prior to the advent of deep learning, multiscale representation, such as scale-invariant feature transform (SIFT) features (Lowe 2004), was used for local feature extraction. It remains valuable for its robustness and generalization in the deep learning era. Methods like FPN (Feature Pyramid Network) (Lin et al. 2017) and PSP (Pyramid Scene Parsing Network) (Zhao et al. 2017) combine convolutional features for object detection and segmentation. Meanwhile, network architectures, e.g., (Chen et al. 2018; Sun et al. 2019; Wang et al. 2019; Huang et al. 2017; Ke, Maire, and Yu 2017), exploit multiscale features effectively. Enhanced designs like OctConv

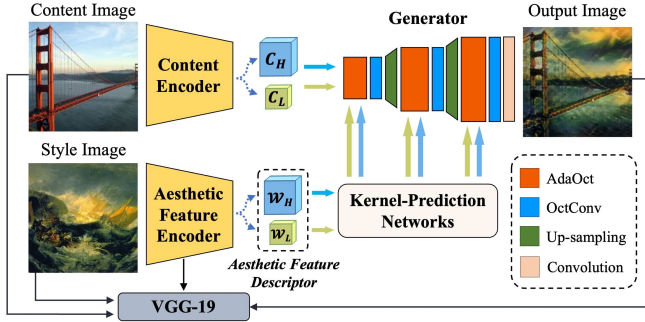


Figure 2: The entire AesFA architecture for aesthetic feature-aware NST. The blue and green arrows indicate the high- and low-frequency feature processes, respectively.

(Chen et al. 2019) exchange inter-frequency information, reducing redundancy and improving CNN classification. Multiscale representation’s prowess is harnessed across various vision tasks: image classification (Wang et al. 2021), compression (Akbari et al. 2020; Liu et al. 2021b), enhancement (Huo, Li, and Zhu 2021; Li et al. 2020; Zhang et al. 2022b), and generation (Wang et al. 2020b; Durall, Pfreundt, and Keuper 2019). Our model, AesFA, decomposes input images by frequencies to extract and transfer aesthetic features from style images, reducing computational costs.

**Frequency analysis in deep learning.** Traditional image processing ((Van Loan 1992; Johnson and Frigo 2006)) has extensively used frequency analysis. Studies connect frequency analysis with deep learning techniques (Chen et al. 2019; Xu et al. 2020, 2019; Durall, Keuper, and Keuper 2020). Wang et al. (Wang et al. 2020a) highlight high-frequency components’ role in neural networks’ generalization. Czoble et al. (Czolbe et al. 2020) introduce a frequency-based reconstruction loss for variational autoencoders (VAEs) using discrete Fourier transformation. Similarly, Cai et al. (Cai et al. 2021) improve identity-preserving image generation by constraining their framework in pixel and Fourier spectral spaces. Nonetheless, these methods are not suitable for NST as perceptual losses are in the feature latent space, not input or output dimensions. We propose contrastive learning for aesthetic features, operating directly in the latent space and significantly reducing computational costs while extracting delicate aesthetic features.

## Method

Here, we introduce a novel methodology that substantially enhances the quality of synthesized images by effectively leveraging the potential of OctConv and decomposing feature maps according to their respective frequencies. The following sections provide a comprehensive examination of the proposed approach and its underlying principles.

### Architecture Overview

As depicted in Figure 2, the AesFA architecture comprises three primary components: a content encoder  $E_c$ , an aesthetic feature encoder  $E_{aes}$  in conjunction with kernel-

prediction networks  $\mathcal{K}$ , and a generator  $G$ . Specifically, the input contents are encoded via a content encoder  $E_c$  and subsequently decomposed into two feature maps containing distinct frequency components. Meanwhile, the style images are processed through the aesthetic feature encoder  $E_{aes}$  to encapsulate higher-level aesthetic feature information. Employing the aesthetic feature descriptor  $\mathcal{W}$ , which is encoded by the aesthetic feature encoder, the kernel-prediction networks  $\mathcal{K}$  predict *aesthetic feature-aware convolutional kernels and biases* for each respective spatial resolution. These predictions then are integrated into the generator alongside the decomposed content latent features. Within the generator, the content features merge with the predicted *aesthetic feature-aware convolutional kernels and biases* for each corresponding frequency using AdaOct. In the terminal layer, the synthesized high- and low-frequency images are amalgamated to produce a single style-transferred output. To summarize, the overarching pipeline proceeds as follows:

1. Encode two decomposed features  $C_H, C_L$  (both the high frequency and low frequency) from the content image  $C$  using the content encoder. To encode the aesthetic feature descriptor  $\mathcal{W}$ , the style image  $S$  is fed to the aesthetic feature encoder  $E_{aes}$ .

$$C_H, C_L := E_c(C), \quad \mathcal{W}_H, \mathcal{W}_L := E_{aes}(S) \quad (1)$$

2. Predict the *aesthetic feature-aware kernels and biases* from the given aesthetic style descriptor  $\mathcal{W}$  using kernel-prediction networks  $\mathcal{K}$ . These kernels and biases will be used in the  $n$ -th layer of the generator.

$$k_{n,H}, b_{n,H} := K_{n,H}(\mathcal{W}_H), \quad k_{n,L}, b_{n,L} := K_{n,L}(\mathcal{W}_L) \quad (2)$$

3. Infuse aesthetic styles with contents in the generator  $G$ , creating style-transferred output  $O$ .

$$O := G(C_H, C_L, \{k_{n,H}, b_{n,H}\}, \{k_{n,L}, b_{n,L}\}) \quad (3)$$

### Frequency Decomposition Networks

**Octave Convolution.** A pivotal aspect of the OctConv operator is its capacity to factorize mixed feature maps by their frequencies while concurrently facilitating efficient communication between high- and low-frequency components. Low-frequency feature maps in OctConv have their spatial resolution reduced by one octave (Lindeberg 2013), where an octave is a spatial dimension divided by a power of two. In this study, a value of 2 was chosen for simplicity. Given input and output of OctConv as  $X = \{X_H, X_L\}$  and  $Y = \{Y_H, Y_L\}$ , the forward pass is defined as:

$$\begin{aligned} Y_H &= f(X_H; W_{H \rightarrow H}) + f(\text{upsample}(X_L, 2); W_{L \rightarrow H}) \\ Y_L &= f(X_L; W_{L \rightarrow L}) + f(\text{pool}(X_H, 2); W_{H \rightarrow L}), \end{aligned} \quad (4)$$

where  $f(X; W)$  represents a convolution with parameters  $W$ . Then,  $\text{pool}(X, 2)$  and  $\text{upsample}(X, 2)$  denote an average pooling operation with kernel size  $2 \times 2$  with a stride of 2 and an upsampling operation by 2 using the nearest interpolation, respectively. Empirical findings indicate that employing OctConv with half the channels for each frequency ( $\alpha = 0.5$ ) yields optimal performance.

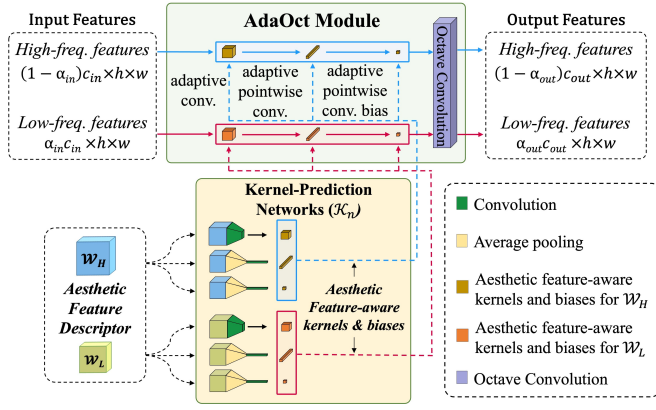


Figure 3: The detailed design of the Adaptive Octave Convolutions (AdaOct) used in AesFA.

**Content and Aesthetic Feature Encoders.** Both proposed encoders improve upon MobileNet (Howard et al. 2017) by replacing all convolutions with OctConv to factorize feature maps by their frequency, thereby reducing network redundancy while maintaining simplicity and efficiency. Spatial reduction in the low-frequency branch expands the receptive field, capturing more contextual information from distant locations and improving performance with greater image resolution. In contrast to the original OctConv, the upsampling order is adjusted to address checkerboard artifacts.

### Aesthetic Feature-Aware Stylization

**Kernel-Prediction Networks.** To effectively apply the aesthetic feature descriptor to content features, we present an approach using kernel-prediction networks similar to those of AdaConv. These networks predict *aesthetic feature-aware kernels and biases* in a depthwise-separable manner, corresponding to frequency and spatial resolution. The *aesthetic feature-aware kernels and biases* comprise depthwise convolution components, pointwise convolution components, and per-channel biases. This approach diverges from the original kernel-prediction network utilized in AdaConv by predicting *aesthetic feature-aware kernels and biases* from both high- and low-frequency aesthetic feature descriptors.

**Adaptive Octave Convolutions.** To efficiently integrate frequency-decomposed contents with the predicted *aesthetic feature-aware kernels and biases*, we begin with AdaConv’s original architecture. However, instead of using it directly, we employ AdaOct followed by an OctConv rather than the standard convolutions outlined in AdaConv. The active interactions between two frequencies that occur in OctConv could further enhance aesthetic stylization quality while reducing the total computational redundancy and unwanted artifacts. Figure 3 provides a detailed overview of our AdaOct module.

The generator comprises three layers, each consisting of an AdaOct module and a standard OctConv block, followed by an upsampling operator. The role of the standard OctConv after the AdaOct module is to learn style-independent

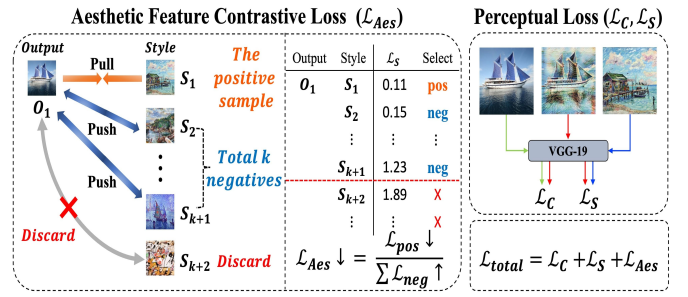


Figure 4: Illustration of the *aesthetic style contrastive loss* in a toy example alongside the other training losses employed in AesFA.

kernels, which aid in the reconstruction of high-quality images. When convolving with *aesthetic feature-aware kernels and biases*, the input channels are grouped into  $n_g$  independent clusters. The network then applies separate spatial and pointwise kernels to learn aesthetic features, such as the microstructure of the texture, as well as cross-channel correlations within the input features. The value of  $n_g$  remains consistent for all *aesthetic feature-aware kernels and biases*, while the remaining parameters adhere to those defined by AdaConv. Notably, AdaConv requires fixed dimensions for style images due to its fully connected layer, while, being fully convolutional, AesFA handles inputs of varying dimensions for both content and style images.

### Aesthetic Feature Contrastive Learning

A singular perceptual loss is insufficient for extracting and expressing intricate aesthetic-style representations (refer to Figure 8). To address this limitation, we adopt and improve the contrastive learning approach from MicroAST (Wang et al. 2023). Our enhanced loss, termed *Aesthetic Feature Contrastive Loss* ( $\mathcal{L}_{Aes}$ ), follows the contrastive learning principle of maintaining proximity between data and their corresponding “positive” samples while distancing them from other instances deemed as “negatives” in the representation space. Consequently, the selection of “positive” and “negative” samples is pivotal for the success of contrastive learning.

Despite its remarkable progress, MicroAST calculates contrastive loss using all negative samples in a mini-batch, making it computationally expensive, particularly at higher resolutions. Intuitively, the nearest sample from the positive offers the most distinctive information. Inspired by hard negative mining techniques (Robinson et al. 2020), we re-defined “negative” samples as a subset of the entire negative sample pool, comprising the  $k$ -th nearest negative samples to the style-transferred output image. For each style-transferred output, the corresponding style image is designated as its positive sample. Meanwhile, the remaining outputs are treated as its *pseudo*-samples. Perceptual losses are defined as the distance between positive and *pseudo*-negative samples, which is calculated by the Exact Feature Distribution Matching (EFDM) algorithm (Zhang et al. 2022a) using the pre-trained VGG-19 network.

The *pseudo*-style perceptual losses are arranged in ascending order, and the top  $k$  *pseudo*-negative samples are selected to compute the final aesthetic feature contrastive loss. The variable  $k$  represents a design choice and can be arbitrarily large, subject to the mini-batch size. In this study, we found that using the nearest style image (i.e.,  $k = 1$ ) yields the best performance. The aesthetic feature contrastive loss is computed at each layer of the entire encoder and on both the high- and low-frequency branches. The formal aesthetic feature contrastive loss is formalized as follows:

$$\mathcal{L}_{Aes} := \sum_{l=1} \mathcal{L}_{Aes,l,High} + \sum_{l=1} \mathcal{L}_{Aes,l,Low}$$

$$\mathcal{L}_{Aes,l} := \sum_{i=1}^N \frac{\|\mathcal{F}_l(O_i) - \text{EFDM}(\mathcal{F}_l(O_i), \mathcal{F}_l(S_{pos,i}))\|_2}{\sum_{j=1}^k \|\mathcal{F}_l(O_i) - \text{EFDM}(\mathcal{F}_l(O_i), \mathcal{F}_l(S_{neg,j}))\|_2}, \quad (5)$$

where  $\mathcal{F}_l(x)$  represents the feature activations of  $l$ -th layer in our encoder given the input  $x$  and  $N$  mini-batch.  $S_{pos}$  and  $S_{neg}$  represent the positive and negative samples for each style-transferred output  $O$ , respectively.

## Training Losses

**Perceptual Loss.** In accordance with previous studies (Gatys, Ecker, and Bethge 2016; Johnson, Alahi, and Fei-Fei 2016), we use the pre-trained VGG-19 model to compute the perceptual loss, which consists of both the content and style losses. However, this work redefines the style loss  $\mathcal{L}_S$  as in the case of EFDM. Given  $I$  representing the stylized output and  $y$  denoting the reference image, the final perceptual losses are as follows:

$$\mathcal{L}_C = \|f_3(I) - f_3(y)\|_2,$$

$$\mathcal{L}_S = \sum_{n=1}^4 \|f_n(I) - \text{EFDM}(f_n(I), f_n(y))\|_2, \quad (6)$$

where  $f_n$  symbolizes the  $n$ -th layer in the VGG-19 model. The content loss is computed at the  $\{conv3\_1\}$  layer in VGG-19, while the style loss is calculated at the  $\{conv1\_1, conv2\_1, conv3\_1, conv4\_1\}$ . It is important to note the VGG-19 model is used solely during training and is entirely excluded from the inference process.

**Total Loss.** Considering all of the aforementioned losses, the total loss is formalized as:

$$\mathcal{L}_{total} = \lambda_C \mathcal{L}_C + \lambda_S \mathcal{L}_S + \lambda_{Aes} \mathcal{L}_{Aes}, \quad (7)$$

where  $\lambda_C$ ,  $\lambda_S$ , and  $\lambda_{Aes}$  are the weighting hyperparameters for each loss. In this paper, we use  $\lambda_C = 1$ ,  $\lambda_S = 10$ , and  $\lambda_{Aes} = 5$ . Figure 8 describes the impact of each hyperparameter.

## Implementation Details

To train our model, we use the COCO dataset (Lin et al. 2014) as content images and the WikiArt dataset (Phillips

and Mackintosh 2011) as style images. During training, images are rescaled to 512 pixels while maintaining the original aspect ratio then randomly cropped to  $256 \times 256$  pixels for augmentation. The model is trained using the Adam optimizer (Kingma and Ba 2014) with a learning rate of 0.0001 and a batch size of 8 for 160,000 iterations. The aesthetic feature has dimensions of (256, 3, 3) for both high- and low-frequency components. All experiments were conducted using the PyTorch framework (Paszke et al. 2019) on a single NVIDIA A100(40G) GPU.

## Experimental Results

In this section, the proposed model’s validity is assessed both qualitatively and quantitatively in comparison to state-of-the-art NST approaches, including AesUST (Wang et al. 2022), AdaIN (Huang and Belongie 2017), AdaConv (Chandran et al. 2021), MicroAST (Wang et al. 2023), EFDM (Zhang et al. 2022a), AdaAttn (Liu et al. 2021a), IECAST (Chen et al. 2021), and StyTr<sup>2</sup> (Deng et al. 2022). We conduct experiments on a range of image resolutions, spanning from small resolutions of 256 pixels to ultra-high 4K resolution. A total of 10 content images and 20 style images are randomly selected for the tests, including images sourced from WikiArt and *pexels.com* for ultra-high resolution images. For each spatial resolution, we generate 200 test results. The results for these methods are acquired by retraining the respective author-released codes using default configurations.

**Qualitative Comparisons.** As described in Figure 5, AesFA qualitatively outperforms eight state-of-the-art NST techniques in terms of aesthetics while maintaining the essential content semantics. AesFA excels in the transfer of unique local aesthetic structural elements from the style image to the content image at all spatial resolutions. Notably, Figure 6 demonstrates that our method can faithfully show aesthetic feature-aware style transfer in terms of tones (first row), texture (second row), brushstrokes (third row), and grains (fourth row). Figure 7 also shows that AesFA excels in transferring the local structure of the style image to the content image in ultra-high resolution (e.g., 4K). MicroAST, for example, suffers from poor aesthetic stylizations and low image quality (blue box in Figure 7). In contrast, AesFA achieves promising outputs with higher image quality (red box in Figure 7). Additional results are demonstrated in the supplementary materials.

**Quantitative Comparisons.** To ensure a comprehensive and effective quantitative comparison, we employ three evaluation metrics: the average SSIM (Wang et al. 2004), the style perceptual loss measured in VGG space (Johnson, Alahi, and Fei-Fei 2016), and the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018). These metrics are used to evaluate the stylization quality in terms of its ability to preserve content and achieve desirable stylization effects. Table 1 shows the quantitative results with various state-of-the-art NST models. Compared to the other techniques, AesFA accomplishes the highest or at least comparable score along all evaluation metrics regardless of image spatial resolution, rendering a single image in less than 0.02 seconds.

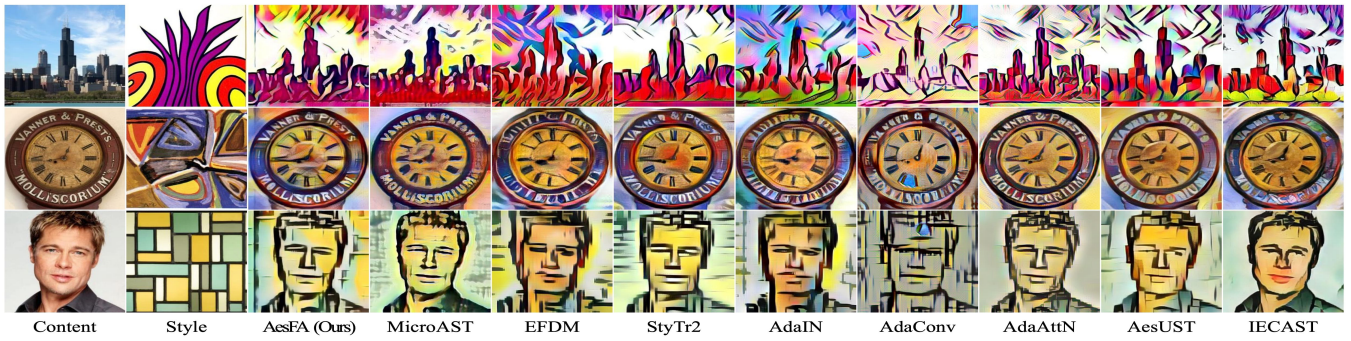


Figure 5: Qualitative comparison with various NST algorithms in 256 pixel resolution. Each column shows the stylized images of different state-of-the-art models.

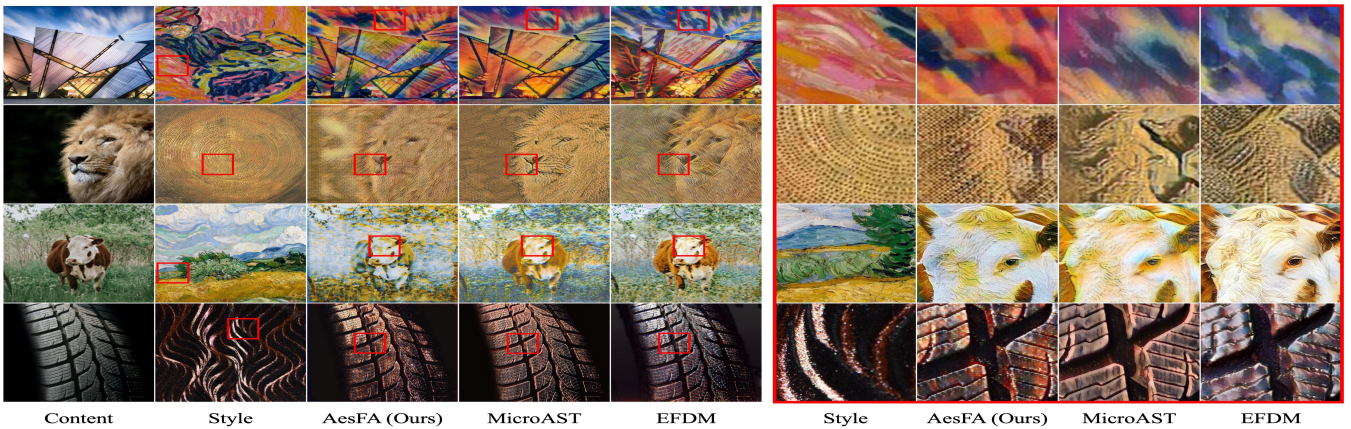


Figure 6: Qualitative comparison with various NST algorithms in 512 (first and second row) and 2K (2048×2048; third and fourth row) resolution. First row: tones, second row: texture, third row: brushstrokes, and fourth row: grains. In all aesthetic features, our AesFA method outperforms.

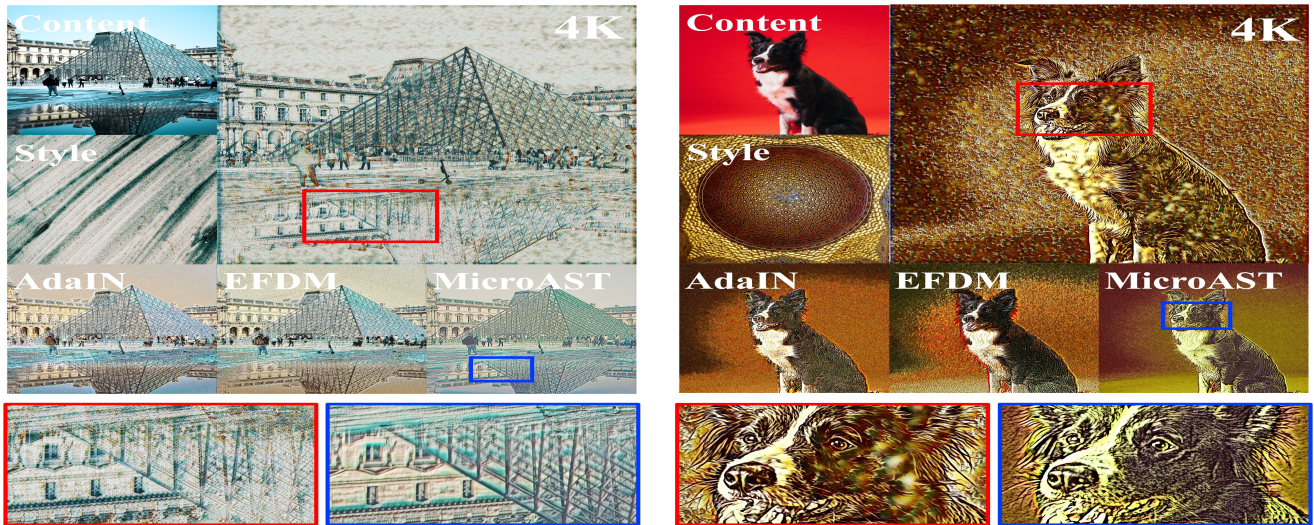


Figure 7: Ultra-high resolution (4K; 4096×4096) comparison. The top-left images are content and style images, and the top right displays our synthesized image. The bottom-left image is the magnified image of our result, and the image on the right is the magnified result from the current state-of-the-art model. Our model outperforms in terms of aesthetic features (e.g., brushstrokes; texture; tones). Zoom in for details.

Resolution	Method	Style Loss ( $\downarrow$ )	LPIPS ( $\downarrow$ )	SSIM ( $\uparrow$ )	Time (sec, $\downarrow$ )	Pref. (% , $\uparrow$ )
256 <sup>2</sup>	AdaConv (2021)	0.936	0.379	0.246	0.493	2.50
	AdaIN (2017)	0.727	0.371	0.230	0.011	11.40
	MicroAST (2023)	1.189	0.372	0.408	<b>0.007</b>	12.35
	EFDM (2022a)	0.720	0.378	0.212	0.011	6.00
	AdaAttn (2021a)	0.993	0.390	<b>0.468</b>	0.027	8.23
	Aes-UST (2022)	0.731	0.372	0.355	0.019	9.18
	IECAST (2021)	0.984	0.392	0.342	0.025	13.28
	StyTr <sup>2</sup> (2022)	<b>0.581</b>	0.377	0.450	0.038	7.90
	<b>AesFA (Ours)</b>	0.692	<b>0.368</b>	0.417	0.016	<b>32.90</b>
	1024 <sup>2</sup>	AdaConv (2021)	N/A	N/A	N/A	N/A
AdaIN (2017)		0.373	0.399	0.336	0.014	4.73
MicroAST (2023)		0.531	0.400	0.430	<b>0.011</b>	15.50
EFDM (2022a)		0.342	0.401	0.313	0.013	6.33
AdaAttn (2021a)		0.596	0.459	0.484	0.060	14.10
Aes-UST (2022)		0.423	0.420	0.455	0.024	14.58
IECAST (2021)		0.554	0.438	0.438	0.015	12.38
StyTr <sup>2</sup> (2022)		0.288	0.411	<b>0.475</b>	1.241	8.88
<b>AesFA (Ours)</b>		<b>0.283</b>	<b>0.392</b>	0.405	0.020	<b>25.03</b>
2048 <sup>2</sup> (2K)		AdaConv (2021)	N/A	N/A	N/A	N/A
	AdaIN (2017)	0.531	0.443	0.311	<b>0.013</b>	19.00
	MicroAST (2023)	0.709	0.447	0.406	0.014	15.18
	EFDM (2022a)	0.475	0.448	0.299	0.018	14.90
	AdaAttn (2021a)	OOM	OOM	OOM	OOM	—
	Aes-UST (2022)	0.754	0.458	<b>0.441</b>	0.028	16.50
	IECAST (2021)	OOM	OOM	OOM	OOM	—
	StyTr <sup>2</sup> (2022)	OOM	OOM	OOM	OOM	—
	<b>AesFA (Ours)</b>	<b>0.404</b>	<b>0.435</b>	0.378	0.020	<b>34.48</b>
	4096 <sup>2</sup> (4K)	AdaConv (2021)	N/A	N/A	N/A	N/A
AdaIN (2017)		0.428	0.376	0.384	0.022	15.53
MicroAST (2023)		0.453	<b>0.371</b>	<b>0.477</b>	<b>0.019</b>	14.88
EFDM (2022a)		0.412	0.379	0.382	0.028	19.00
AdaAttn (2021a)		OOM	OOM	OOM	OOM	—
Aes-UST (2022)		OOM	OOM	OOM	OOM	—
IECAST (2021)		OOM	OOM	OOM	OOM	—
StyTr <sup>2</sup> (2022)		OOM	OOM	OOM	OOM	—
<b>AesFA (Ours)</b>		<b>0.216</b>	0.373	0.469	<b>0.020</b>	<b>50.60</b>

Table 1: Quantitative comparison with various state-of-the-art NST algorithms. “N/A” means “Not applicable at this resolution” and “OOM” stands for “Out of GPU memory”.

**User Study.** Evaluating the outcomes of stylization is a highly subjective matter. Hence, we have conducted a user study for the nine approaches. We randomly show each participant 20 ballots (4 ballots for each resolution) containing the content, style, and nine outputs. For each ballot, participants were given unlimited time to select their favorite output in terms of aesthetically pleasing stylization and content preservation. We collected 1,580 valid votes from 79 subjects. The preference percentage of each method for each resolution is included in the last column of Table 1. The user study results demonstrate that our stylized images are more appealing than or at least comparable to the competitors.

### Ablation Studies

We also have conducted a series of ablation studies to provide justification for the architectural decisions employed and to highlight their effectiveness. We first explore the effect of *aesthetic feature contrastive loss*,  $\mathcal{L}_{Aes}$  in Figure 8. When training without  $\mathcal{L}_{Aes}$ , the stylization quality of the proposed model drastically degrades, and unsatisfactory artifacts appear (e.g., the stripe pattern in the background). This shows that the newly devised loss revealed by AesFA plays an important role in expressing aesthetic features and eliminating artifacts. Notably, we changed the alpha value

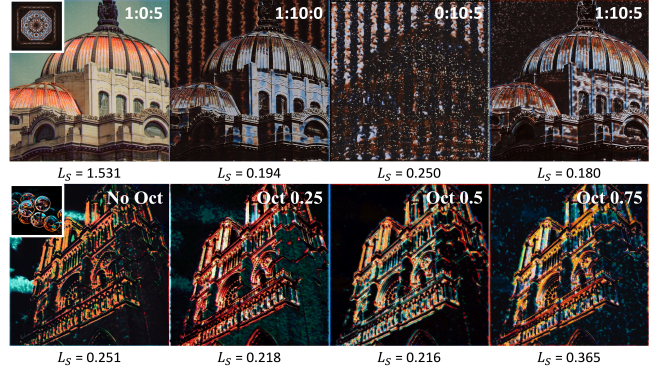


Figure 8: Top: The effectiveness of the proposed *aesthetic feature contrastive loss*. Bottom: The effectiveness of the  $\alpha$  value of OctConv.  $\mathcal{L}_S$  denotes the style perceptual loss.

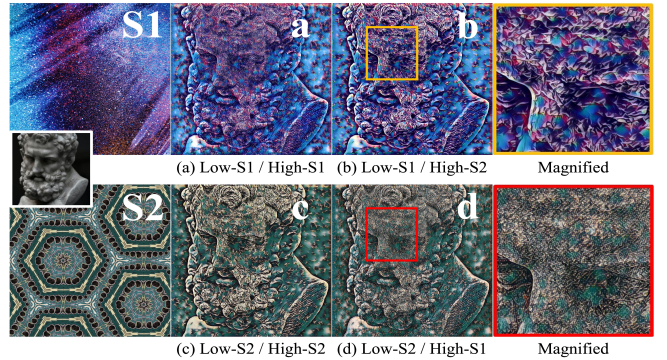


Figure 9: The style blending results generated by AesFA with different style images. Zoom in for details.

( $\alpha$ ) of the OctConv, which denotes the ratio of the number of low-frequency channels to the total-frequency channels. Results show that our model with  $\alpha = 0.5$  performs the best qualitatively (Figure 8) and quantitatively, significantly reducing artifacts in the background and enhancing stylization quality. Meanwhile, the model with standard convolutions (No Oct) shows undesirable artifacts in the background and poor quality of colorization. Detailed descriptions and images of high- and low-frequency component images for each setting are provided in the supplementary materials.

Figure 9 shows the style blending, i.e., using the low- and high-frequency style information from different style images. Sub-figures (a)-(d) show different combinations of origins for low- and high-frequency style information. For instance, “(b) Low-S1 / High-S2” indicates that we use the low-frequency style information from image “S1” and the high-frequency style information from image “S2”.

### Conclusion

In this study, we propose AesFA, a lightweight and effective model for aesthetic feature-aware NST. Unlike existing models, AesFA decomposes the image by frequencies and infuses it with corresponding aesthetic features. We intro-

duce a new aesthetic feature contrastive loss by leveraging pretrained VGGs to guide stylization effectively. Our experiments demonstrate that the model and new loss significantly enhance the quality of generated images *regardless of resolution*. Furthermore, AesFA achieves stylized output in less than 0.02 seconds, making it suitable for real-time ultra-high resolution rendering (4K) applications.

### Acknowledgements

This work was supported by the U.S. Department of Energy (DOE), Office of Science (SC), Advanced Scientific Computing Research program under award DE-SC-0012704 and used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 using NERSC award ASCR-ERCAP0023081.

This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C1006503, 2021K1A3A1A2103751212, 2021M3E5D2A01022515, RS-2023-00266787, RS-2023-00265406), by Creative-Pioneering Researchers Program through Seoul National University(No. 200-20230058), by Semi-Supervised Learning Research Grant by SAMSUNG(No.A0426-20220118), and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

### References

- Akbari, M.; Liang, J.; Han, J.; and Tu, C. 2020. Generalized octave convolutions for learned multi-frequency image compression. *arXiv preprint arXiv:2002.10032*.
- An, J.; Li, T.; Huang, H.; Ma, J.; and Luo, J. 2023. Is Bigger Always Better? An Empirical Study on Efficient Architectures for Style Transfer and Beyond. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4084–4094.
- Cai, M.; Zhang, H.; Huang, H.; Geng, Q.; Li, Y.; and Huang, G. 2021. Frequency domain image translation: More photo-realistic, better identity-preserving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13930–13940.
- Chandran, P.; Zoss, G.; Gotardo, P.; Gross, M.; and Bradley, D. 2021. Adaptive convolutions for structure-aware style transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7972–7981.
- Chen, C.-F.; Fan, Q.; Mallinar, N.; Sercu, T.; and Feris, R. 2018. Big-little net: An efficient multi-scale feature representation for visual and speech recognition. *arXiv preprint arXiv:1807.03848*.
- Chen, H.; Wang, Z.; Zhang, H.; Zuo, Z.; Li, A.; Xing, W.; Lu, D.; et al. 2021. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems*, 34: 26561–26573.
- Chen, T. Q.; and Schmidt, M. 2016. Fast patch-based style transfer of arbitrary style. *arXiv preprint arXiv:1612.04337*.
- Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; and Feng, J. 2019. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3435–3444.
- Czolbe, S.; Krause, O.; Cox, I.; and Igel, C. 2020. A loss function for generative neural networks based on watson’s perceptual model. *Advances in Neural Information Processing Systems*, 33: 2051–2061.
- Deng, Y.; Tang, F.; Dong, W.; Ma, C.; Pan, X.; Wang, L.; and Xu, C. 2022. StyTr<sup>2</sup>: Image Style Transfer with Transformers. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Dumoulin, V.; Shlens, J.; and Kudlur, M. 2016. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*.
- Durall, R.; Keuper, M.; and Keuper, J. 2020. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7890–7899.
- Durall, R.; Pfreundt, F.-J.; and Keuper, J. 2019. Stabilizing GANs with Soft Octave Convolutions. *arXiv preprint arXiv:1905.12534*.
- Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423.
- Gatys, L. A.; Ecker, A. S.; Bethge, M.; Hertzmann, A.; and Shechtman, E. 2017. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3985–3993.
- Gentleman, W. M.; and Sande, G. 1966. Fast Fourier transforms: for fun and profit. In *Proceedings of the November 7-10, 1966, fall joint computer conference*, 563–578.
- Ghiasi, G.; Lee, H.; Kudlur, M.; Dumoulin, V.; and Shlens, J. 2017. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv preprint arXiv:1705.06830*.
- Howard, A. G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; and Adam, H. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Huang, G.; Chen, D.; Li, T.; Wu, F.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844*.
- Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.
- Huo, F.; Li, B.; and Zhu, X. 2021. Efficient wavelet boost learning-based multi-stage progressive refinement network for underwater image enhancement. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, 1944–1952.
- Jing, Y.; Liu, X.; Ding, Y.; Wang, X.; Ding, E.; Song, M.; and Wen, S. 2020. Dynamic Instance Normalization for Arbitrary Style Transfer. In *AAAI*.
- Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, 694–711. Springer.
- Johnson, S. G.; and Frigo, M. 2006. A modified split-radix FFT with fewer arithmetic operations. *IEEE Transactions on Signal Processing*, 55(1): 111–119.
- Ke, T.-W.; Maire, M.; and Yu, S. X. 2017. Multigrid neural architectures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6665–6673.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kotovenko, D.; Sanakoyeu, A.; Lang, S.; and Ommer, B. 2019. Content and style disentanglement for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4422–4431.
- Li, S.; Cai, Q.; Li, H.; Cao, J.; Wang, L.; and Li, Z. 2020. Frequency separation network for image super-resolution. *IEEE Access*, 8: 33768–33777.
- Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; and Yang, M.-H. 2017. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30.
- Lin, T.; Ma, Z.; Li, F.; He, D.; Li, X.; Ding, E.; Wang, N.; Li, J.; and Gao, X. 2021. Drafting and revision: Laplacian pyramid network for fast high-quality artistic style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5141–5150.
- Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, 740–755. Springer.
- Lindeberg, T. 2013. *Scale-space theory in computer vision*, volume 256. Springer Science & Business Media.
- Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; and Ding, E. 2021a. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6649–6658.
- Liu, Z.; Meng, L.; Tan, Y.; Zhang, J.; and Zhang, H. 2021b. Image compression based on octave convolution and semantic segmentation. *Knowledge-Based Systems*, 228: 107254.
- Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60: 91–110.
- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Phillips, F.; and Mackintosh, B. 2011. Wiki Art Gallery, Inc.: A case for critical thinking. *Issues in Accounting Education*, 26(3): 593–608.
- Robinson, J.; Chuang, C.-Y.; Sra, S.; and Jegelka, S. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.
- Shen, F.; Yan, S.; and Zeng, G. 2018. Neural style transfer via meta networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8061–8069.
- Sheng, L.; Lin, Z.; Shao, J.; and Wang, X. 2018. Avatar-net: Multi-scale zero-shot style transfer by feature decoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8242–8250.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5693–5703.
- Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.
- Ulyanov, D.; Lebedev, V.; Vedaldi, A.; and Lempitsky, V. 2016. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*.
- Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6924–6932.
- Van Loan, C. 1992. *Computational frameworks for the fast Fourier transform*. SIAM.
- Wang, H.; Kembhavi, A.; Farhadi, A.; Yuille, A. L.; and Rastegari, M. 2019. Elastic: Improving cnns with dynamic scaling policies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2258–2267.
- Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020a. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8684–8694.
- Wang, J.; Guo, S.; Huang, R.; Li, L.; Zhang, X.; and Jiao, L. 2021. Dual-channel capsule generation adversarial network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–16.

- Wang, P.; Li, Y.; and Vasconcelos, N. 2021. Rethinking and improving the robustness of image style transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 124–133.
- Wang, Y.; Khan, S.; Gonzalez-Garcia, A.; Weijer, J. v. d.; and Khan, F. S. 2020b. Semi-supervised learning for few-shot image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4453–4462.
- Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.
- Wang, Z.; Zhang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Xing, W.; and Lu, D. 2022. AesUST: towards aesthetic-enhanced universal style transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1095–1106.
- Wang, Z.; Zhao, L.; Zuo, Z.; Li, A.; Chen, H.; Xing, W.; and Lu, D. 2023. MicroAST: Towards Super-Fast Ultra-Resolution Arbitrary Style Transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Xu, K.; Qin, M.; Sun, F.; Wang, Y.; Chen, Y.-K.; and Ren, F. 2020. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1740–1749.
- Xu, Z.-Q. J.; Zhang, Y.; Luo, T.; Xiao, Y.; and Ma, Z. 2019. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint arXiv:1901.06523*.
- Yoo, J.; Uh, Y.; Chun, S.; Kang, B.; and Ha, J.-W. 2019. Photorealistic style transfer via wavelet transforms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9036–9045.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.
- Zhang, Y.; Li, M.; Li, R.; Jia, K.; and Zhang, L. 2022a. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8035–8045.
- Zhang, Y.; Li, Q.; Qi, M.; Liu, D.; Kong, J.; and Wang, J. 2022b. Multi-scale frequency separation network for image deblurring. *arXiv preprint arXiv:2206.00798*.
- Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.