We thank all reviewers for their constructive comments to which we respond below.

To R2: About the differences between the MicroAST and the proposed method. Our work differs from MicroAST in three aspects: 1) AesFA processes images at different spatial frequencies to better extract aesthetic features, whereas MicroAST relies on standard convolution. 2) AesFA predicts "aesthetic feature-aware kernels and biases" in a depthwise-separable manner, preserving spatial characteristics and facilitating modulation based on style, including correlations across input channels and frequencies. In contrast, MicroAST relies on summary statistics, which may not adequately capture aesthetic features. 3) Inspired by hard negative mining, we propose a new contrastive learning method for aesthetic features, using the k-th nearest negative samples to the stylized output. In contrast, MicroAST employs all negative samples in a mini-batch, leading to inefficient training, particularly with ultra-high resolutions.

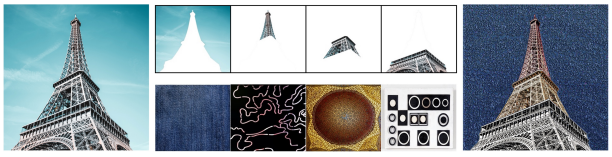

Figure 1: The granular control of aesthetic style attributes in 4K (4096) resolution. Magnify the image to see the details.

To R3: About the granular control over the defined style attributes. Spatial control is also coveted by users who wish to modify an image by applying different styles to various regions of the image. Fig. 1 shows an example of spatially controlling the stylization in 4K resolution using AesFA. A set of masks is additionally required. We explore finer aesthetic style control by encoding style images into various frequencies, enabling users to blend content and style without additional resources. For example, in Fig. 9 in the main text, we show the style blending, i.e., using the low-frequency and high-frequency style information from different images. We can see that the style transferred images keep the color information from the low-frequency image and change the texture information based on the high-frequency image. This could also be successfully adapted to the video style transfer. Additional result images are in supplementary materials.

To R5: About the motivation behind the frequency division processing of images and its effectiveness. Visual information is conveyed across various frequencies, with higher frequencies carrying fine details like texture and edges and lower frequencies encoding global structures such as colors. Our results show that disentangling information across different frequencies indeed helps extract richer visual information. It also enhances the stylization by transferring different aesthetic features to its corresponding frequency band. The effectiveness of this frequency division strategy is illustrated in Tab. 1, Tab. 2, and Tab. 3 of supplementary materials.

To R5: About the difference between our proposed method and the WCT2 (Yoo et al. 2019) and LapStyle (Lin et al. 2021). Thanks for pointing out the related works. The primary goal of WCT2 is to achieve photorealistic style transfer. To achieve photorealism, a model must apply the style to the content while preserving the intricate details of the image. Therefore, WCT2 aims to retain high-frequency information to the greatest extent possible. However, the 'aesthetic style' encompasses both structural information and color information. For instance, in Fig.2 (a), style can be expressed not only by color but also by structural elements such as swirling thick lines, and these aesthetic characteristics are conveyed through different frequencies within the image. Thus, transferring it to the contents shall accompany both changes in high- and low-frequency features. Similarly, LapStyle aims to the artistic style transfer, but it requires an additional discriminative network, and they do not use a frequency division strategy to encode the style but rather use it to generate stylized images using the style features encoded by a pre-trained VGG network. We evaluated LapStyle in 256 resolution and get: VGG style loss of 1.412, LPIPS of 0.392, and inference time of 0.030s (the lower the better; Ours: **0.692**, **0.368**, and **0.016s**, respectively).

Figure 2: The qualitative comparison between ours and LapStyle.

To R5: Capturing spatial information related to style representations. Departing from the previous approaches (e.g., AdaConv) that utilize fully connected layers at the end to compute the spatially invariant style based on the input features, we chose to omit the fully connected layers to retain the spatial information related to aesthetic characteristics. Furthermore, previous approaches transfer a simple pair of global statistics (e.g., mean, variance) from the style. Our proposed module, AdaOct, effectively transfers 3D aesthetic kernels and biases that encode richer structural and statistical characteristics to the contents. By employing this approach, our model maintains the integrity of spatial information, while also establishing correlations among features across different input channels and frequencies.

To R5: Inference time. When dealing with higher resolutions, the inference time of AesFA marginally increases.

Image resolutions	256^2	512^2	1024^2	$2048^2(2k)$	$4096^2(4k)$
Inference time	0.01559	0.01725	0.01951	0.01972	0.02033

Table 1: Inference time for AesFA with different test image resolutions. The unit is sec/image.

To R5: The detailed explanations of Figure 9 in the main text. 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-1 / High-2" indicates that we use the low-frequency style information from image "1" and the high-frequency style information from image "2".

To R5: Notation. Thank you for pointing out these issues. We thoroughly checked and revised the manuscript.