Volume xx (20yy), Number z, pp. 1–13

Preserving Photographic Defocus in Stylised Image Synthesis

Hong-Yi Wang [1] and Yu-Ting Wu [†2]

[1] s711283120@gm.ntpu.edu.tw, National Taipei University, Taiwan
[2] yutingwu@mail.ntpu.edu.tw, National Taipei University, Taiwan

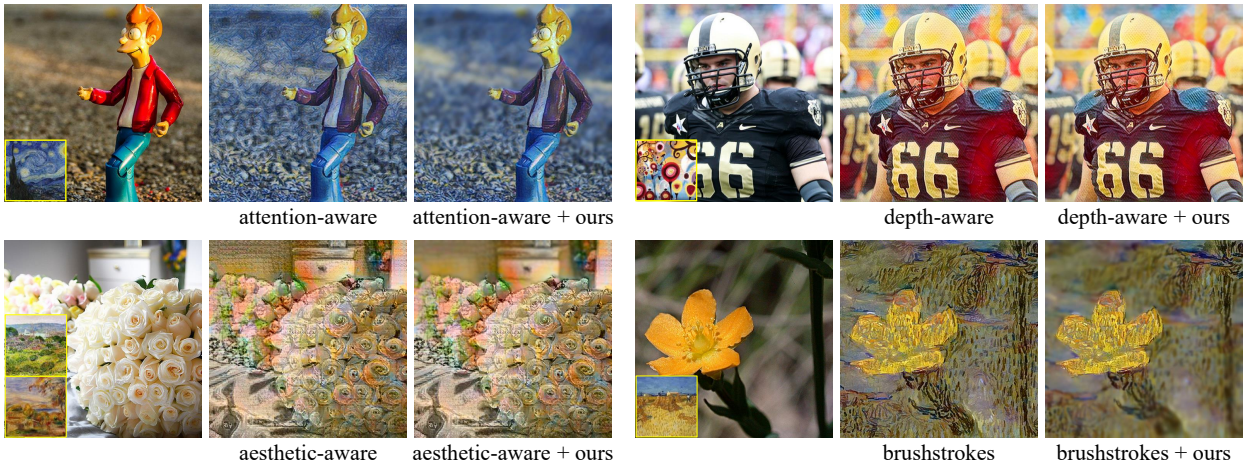

Figure 1: By integrating our method with existing style transfer techniques, including attention-aware [ZXW*24], depth-aware [IM22], aesthetic-aware [HJL*20], and parameterized brushstrokes-based [KWHO21] methods, we effectively preserve the original defocus effects intended by the photographer. The incorporation of defocus into the stylized outputs further enhances structural coherence, allowing the background to fade naturally and improving visual separation and emphasis on focal regions. For each method, the corresponding content and style images are presented in the first column.

Abstract

While style transfer has been extensively studied, most existing approaches fail to account for the defocus effects inherent in content images, thereby compromising the photographer's intended focus cues. To overcome this shortcoming, we introduce an optimisation-based post-processing framework that restores defocus characteristics to stylised images, regardless of the style transfer technique used. Our method initiates by estimating a blur map through a data-driven model that predicts pixel-level blur magnitudes. This blur map subsequently guides a layer-based defocus rendering framework, which effectively simulates depth-of-field (DoF) effects using a Gaussian filter bank. To map the blur values to appropriate kernel sizes in the filter bank, we introduce a neural network that determines the optimal maximum filter size, ensuring both content integrity and stylistic fidelity. Experimental results, both quantitative and qualitative, show that our method significantly improves stylised images by preserving the original depth cues and defocus details.

CCS Concepts

• ***Computing methodologies*** → ***Image processing; Computational photography; Non-photorealistic rendering;***

1. Introduction

Artificial intelligence (AI) has profoundly transformed various domains by enabling intelligent data processing and adaptive learning. Among its many applications, neural style transfer stands out as a unique intersection of artistic expression and technological

† Corresponding author

advancement. This technique involves transferring the stylistic attributes or textures of a reference image onto a target one, producing a synthesized image that preserves the original content while reflecting the aesthetic qualities of the chosen style. Modern digital illustration and image editing tools are increasingly integrating such techniques to enrich user experiences with creative and entertainment-driven features.

Recent progress in style transfer techniques has increasingly focused on enhancing the preservation of structural integrity in content images while altering their stylistic attributes. Certain approaches incorporate semantic information to better retain object boundaries and shapes [LH22, LZ22, ZDCY23], whereas others adapt stroke patterns based on surface depth cues [LCLR17, CLW*20, IM22, CZLY23] or leverage self-attention mechanisms for more context-aware transformations [YRX*19, PL19, DTD*22, ZHW*23]. Despite these advancements, most existing methods overlook defocus effects, an important photographic principle that is either deliberately applied to highlight focal subjects or naturally results from lens limitations. Prior studies [KTJ06, KSL*16], based on interviews with both professional and amateur photographers, have highlighted key traits of high-quality photographs. Among these, a defining characteristic is the contrast between a sharply focused foreground and a blurred background, typical of a shallow depth of field. Consequently, maintaining or recovering defocus effects in stylized images is crucial.

To the best of our knowledge, the method proposed by Cao *et al*. [CCJT23] remains the only existing approach that explicitly integrates defocus effects into the style transfer pipeline. Their technique utilizes a stroke pyramid to assign varying stroke sizes based on the level of defocus. Although this multi-scale stroke mechanism enhances content readability, it does not faithfully reproduce the blurred characteristics inherent to the original image. Additionally, the method's effectiveness degrades when applied to style images with complex or highly detailed structures.

To address the limitations of prior approaches, this paper proposes transferring defocus characteristics from the original content image to the stylized output. A straightforward approach for achieving this involves applying blur filters whose kernels are guided by the blur distribution in the original image. Ideally, these filters should replicate the defocus while preserving edge integrity and structural details. However, adjusting parameters for edge-preserving filters, such as bilateral filters [TM98], is both challenging and computationally demanding, particularly when dealing with spatially varying kernel sizes. To address this challenge, we adopt a layer-based defocus rendering framework, which enables efficient simulation of depth-of-field (DoF) effects through a Gaussian filter bank while preserving edges and structural fidelity.

Our method begins by utilizing a data-driven defocus estimation model [LLCL19] to generate a pixel-wise blur map of the content image. Next, we introduce a kernel mapping network that determines the optimal maximum filter size for constructing a Gaussian filter bank. This optimization is guided by two objectives: preserving the perceptual integrity of the content image, measured using the Learned Perceptual Image Patch Similarity (LPIPS) metric [ZIE*18], and ensuring stylistic coherence with the reference image via a texture-based loss. Finally, the constructed filter bank is used in a layer-based defocus rendering algorithm to reproduce realistic defocus effects in the stylized output.

A notable advantage of our method lies in its post-processing nature, making it inherently model-agnostic and readily compatible with existing style transfer techniques to enhance results by reintroducing defocus effects. As demonstrated in Figure 1, our method can be integrated with various style transfer techniques, including attention-aware [ZXW*24], depth-aware [IM22], aesthetic-aware [HJL*20], and parameterized brushstrokes-based [KWHO21] methods. Once applied, the stylized outputs successfully preserve the original defocus characteristics. Additionally, structural fidelity in the foreground is enhanced due to perceptually coherent background attenuation introduced by the restored DoF effects.

2. Related Work

2.1. Neural style transfer

Gatys *et al*. [GEB16] proposed the pioneering neural style transfer (NST) algorithm, which leverages convolutional neural networks (CNNs) to extract features from both content and style images. By minimizing the differences between these features, their method generates a stylized image that preserves the structural content of the input while embedding the visual characteristics of the target style. Since then, numerous studies have aimed to advance NST, focusing on improvements in computational efficiency, generalization, structural fidelity, visual quality, and controllability.

Computational efficiency. The method proposed by Gatys *et al*. [GEB16] employs iterative optimization to generate the stylized images, which is highly demanding in terms of runtime and memory. To address these limitations, later works introduced more efficient feed-forward network architectures [JAFF16, ULVL16], enabling stylized image generation in a single forward pass.

Generalization. To eliminate the need for training a separate model for each style, numerous approaches have been developed that allow a single model to handle multiple styles [ZD18, LFY*17a] or even arbitrary ones [CS16, HB17, LFY*17b, WSZ*20, DTD*20, WHSX21, DTD*21, LLW*23]. This generalization has been achieved through diverse strategies, such as local patch matching between content and style features [CS16], aligning the statistical moments of content and style features [HB17], applying whitening and coloring transforms within an image reconstruction framework [LFY*17b], leveraging transformer-based composition via a style bank [WHSX21], and disentangling and recombining content and style representations [DTD*20, LLW*23].

Structural fidelity. The structural details of the content image, such as object shapes, edges, and spatial layout, are often diminished in the stylized output. To improve structural fidelity, several approaches have been proposed, including semantic alignment between content and style images [LH22], the introduction of structure loss [LZ22] and edge-aware loss [ZDCY23], the integration of depth estimation modules to better retain object boundaries across depth layers [LCLR17, CLW*20, IM22], and the adoption of self-attention mechanisms [YRX*19, PL19, ZHW*23, ZXW*24].

Visual quality. To narrow the gap between real artworks and synthesized stylizations, several methods [CZW*21a, CZW*21b] have

been developed that learn global tone or overall stylistic characteristics from a collection of artworks, while deriving fine-grained attributes such as color and texture from a specific style image. To further enhance stylization quality, new loss functions have been introduced, including attention-guided content loss for improved preservation of semantic relationships, patch-based style loss to enhance local style consistency [CZLY23], and style-aware normalized loss to mitigate biases inherent in traditional style losses [CJW*21]. In addition, Hu *et al*. [HJL*20] demonstrated that aesthetic attributes can be decomposed into color and texture components, and proposed a method to independently transfer these components using separate reference images.

Controllability. Controllability in style transfer can be broadly categorized into multi-domain and multimodal approaches. Multi-domain style transfer enables controlled output generation either through user-specified parameters [JLY*18, ZZL*22, TLL*23, HJL*23] or by incorporating guidance from multiple stylistic domains [CWC20, HJY*21]. In contrast, multimodal style transfer enhances output diversity by producing multiple distinct stylizations from a single reference style [CZZ*21, CWJ*23]. Lin *et al*. [LTD*21] introduced a unified framework that supports both paradigms simultaneously through a style distribution alignment module.

For alternative objectives, certain approaches prioritize brush stroke synthesis over direct pixel-level generation [KWHO21, WLZF22, ZLZ23, LWH24]. StylerDALLE [XSS23] leverages CLIP-based language supervision to eliminate the need for extensive data collection and the manual design of specialized loss functions. Beyond artistic stylization, style transfer has also been applied in other domains, for instance, to reduce domain bias in monocular depth estimation using synthetic data [AAB18] and to generate fraudulent face images for face liveness detection [RMN*19].

Aligned with our objectives, Cao *et al*. [CCJT23] incorporate defocus cues from the content image into the style transfer process by modulating stroke sizes according to defocus levels using a stroke pyramid architecture. In contrast, our method reconstructs the defocused appearance by determining spatially varying blur kernel size, enabling more precise alignment with the original content image and improved separation between foreground and background regions. Additionally, as a post-processing framework, our method is agnostic to the underlying style transfer algorithm, allowing it to be seamlessly integrated with both existing and future style transfer techniques.

2.2. Defocus rendering

Defocus arises because a camera lens can sharply focus only on objects at a specific distance, known as the focal distance. To simulate depth of field (DoF), many approaches have been developed that take depth information as input. Such depth data can be obtained either from hardware-based sensing [YLY*16, WGJ*18] or through computational depth estimation techniques [YTy*11, SWS*17, SGW*18, WSZ*18, PCL*22, PCL*25]. To model defocus effects, Yang *et al*. [YLY*16] introduced a depth-aware pseudo ray-tracing approach that approximates light field rendering while improving memory and computational efficiency. Wadhwa *et al*. [WGJ*18] generated shallow DoF effects in mobile photography by leveraging person segmentation masks or dual-pixel autofocus data. Srinivasan *et al*. [SGW*18] trained a depth estimation network using images captured under varying aperture settings, whereas Wang *et al*. [WSZ*18] relied on RGB-D datasets for depth acquisition. SteReFo [BHMS19] proposed a trainable framework that jointly optimizes stereo depth estimation and refocusing, with user-controllable parameters such as the focus plane and aperture. Their method discretizes depth into multiple layers, applying independent blurring to each, which improves defocus simulation efficiency while also reducing color bleeding from filtering.

Defocus effects can also be synthesized by reconstructing light-field representations, either through learning-based techniques [SWS*17] or geometric warping-based methods [YTy*11]. More recently, Seizinger *et al*. [SCK*23] proposed a data-driven framework capable of transferring defocus and bokeh characteristics across different camera lens configurations. BokehMe and its subsequent extension [PCL*22, PCL*25] achieve high-quality bokeh rendering by combining classical techniques with data-driven refinement, employing a neural renderer to correct artifacts introduced by the classical pipeline.

2.3. Defocus blur detection and estimation

To restore defocus effects in the stylized output, our method estimates the defocus level for each pixel of the original content image, a task closely related to defocus blur detection and estimation. Traditional approaches typically rely on low-level, hand-crafted features such as gradients [ZS11, XQJ17, KJ18], contrast [YE16], or frequency domain information [SXJ14, TWH*16]. However, these approaches often struggle in homogeneous regions where edge information is limited. With the rise of deep learning, high-level semantic features have been increasingly exploited [KSP*18, LLCL19, CP20, ZSL21, ZLZY22], leading to substantial gains in both robustness and accuracy.

Defocus map estimation from blurred images supports a wide range of applications, including image editing [BD07], image quality assessment [SBC12], saliency detection [JLYP13], depth estimation [ZS11], and image deblurring [RCLL22, ZHW*24]. In fact, some approaches explicitly integrate defocus blur detection or estimation into their pipelines, for example, the deblurring approach introduced by Zhao *et al*. [ZHW*24].

3. Method

3.1. Overview

Figure 2 presents an overview of the proposed method. First, an arbitrary style transfer model processes the content and style images to generate the stylized output. In parallel, the content image is fed into a blur estimation network [LLCL19] to produce a corresponding blur map. This blur map then guides the subsequent process of transferring defocus characteristics from the original content image to the stylized output.

In contrast to most defocus rendering techniques, which emphasize controllable DoF effects or refocusing (see Sec. 2.2), our

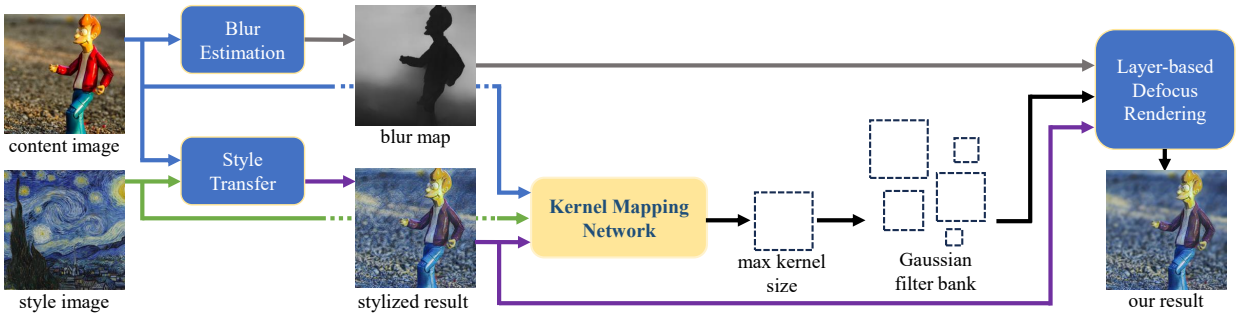

Figure 2: **Flowchart of our method.** Our approach enhances the output of an arbitrary style transfer model by reintroducing the defocus characteristics of the original content image into the stylized result. Initially, the content and style images are processed by a selected style transfer technique to generate a stylized output. Simultaneously, the content image is passed through a blur estimation model [LLCL19] to generate a blur map that captures pixel-wise defocus intensity. The content, style, and stylized images are then input into our Kernel Mapping Network, which predicts the optimal maximum blur kernel size for constructing a Gaussian filter bank. This filter bank, together with the blur map, is employed within a layer-based defocus rendering framework [KS07, BHMS19] to accurately reproduce depth-of-field (DoF) effects in the final output.

objective is to preserve and replicate the defocus effects already present in the content image. Instead of simulating the physical process of defocus, we directly predict the resulting blur. This design simplifies the overall pipeline and eliminates the need for depth estimation or manual specification of focal distance, a parameter that is both challenging for users to provide reliably and typically ambiguous, since estimated depth maps often encode relative rather than absolute depth.

Our method produces defocus effects by adaptively blurring the stylized image, with a strong emphasis on preserving edges and structural details. Although edge-preserving filters such as bilateral filters [TM98] offer a direct solution, selecting appropriate spatial and range parameters simultaneously at the pixel level is challenging. To address this, we adopt the layer-based defocus rendering framework proposed by [BHMS19], which preserves edge integrity using Gaussian filters. This framework discretizes the scene into multiple blur levels, each corresponding to a specific filter size. Together, collectively forming a Gaussian filter bank.

The Gaussian filter bank is constructed by assigning a specific filter size to each blur level. To establish an optimal mapping, we introduce a neural network, referred to as the *Kernel Mapping Network*, which takes the content image, the style image, and the stylized result as inputs. This network predicts the optimal maximum filter size that effectively captures the defocus characteristics of the original content image while preserving stylistic coherence. Once the maximum filter size is determined, the remaining filter sizes in the filter bank are derived proportionally. Gaussian filtering is then applied at each blur level using its corresponding filter size on the associated pixels, producing a stylized image that seamlessly integrates spatially varying blur with the intended stylization.

The remainder of this section details the algorithms behind each pipeline component. Sec. 3.2 introduces the blur estimation model used to quantify the blur intensity in the content image. Sec. 3.3 details the architecture and loss functions of the kernel mapping network, which predicts the optimal maximum kernel size for building the Gaussian filter bank. Lastly, Sec. 3.4 describes the layer-based defocus rendering framework employed to replicate the original content image's defocus effects using the constructed filter bank.

3.2. Blur estimation

Our method employs the pretrained DMENet model [LLCL19], an end-to-end convolutional neural network specifically designed for pixel-wise defocus estimation in blurred images. We chose this model for its consistently reliable performance on real-world photographs observed in our experiments. Its robustness can be primarily attributed to its training on a substantial dataset of synthetic images, where blur is simulated using image–depth map pairs. Moreover, a domain adaptation strategy is employed to narrow the gap between synthetic and real-world data. Nonetheless, alternative methods for defocus blur estimation could also be adopted.

3.3. Kernel mapping network and optimization scheme

The kernel mapping network, depicted in Figure 3, is designed to determine the maximum filter size K corresponding to the highest level of blur in the blur map. To achieve this, we concatenate the content image, style image, and stylized image along the channel dimension, forming a 9-channel input tensor. This combination enables the network to jointly consider content structure, stylistic attributes, and stylization effects when estimating the optimal blur level. To extract meaningful features, we employ a ResNet-18 backbone pretrained on ImageNet [DDS*09], leveraging its strong semantic representation capabilities. Prior to feeding the input into ResNet-18, we reduce the 9-channel tensor to the standard 3-channel format using a 3×3 convolution followed by batch normalization. The extracted features are subsequently processed through an additional convolutional layer, followed by global average pooling, layer normalization, and a fully connected layer. In practice, we observed that the output from the fully connected layer can be unstable. To address this, we apply a softplus activation function to produce more stable and reliable predictions.

In each iteration of optimization, the network predicts a max-

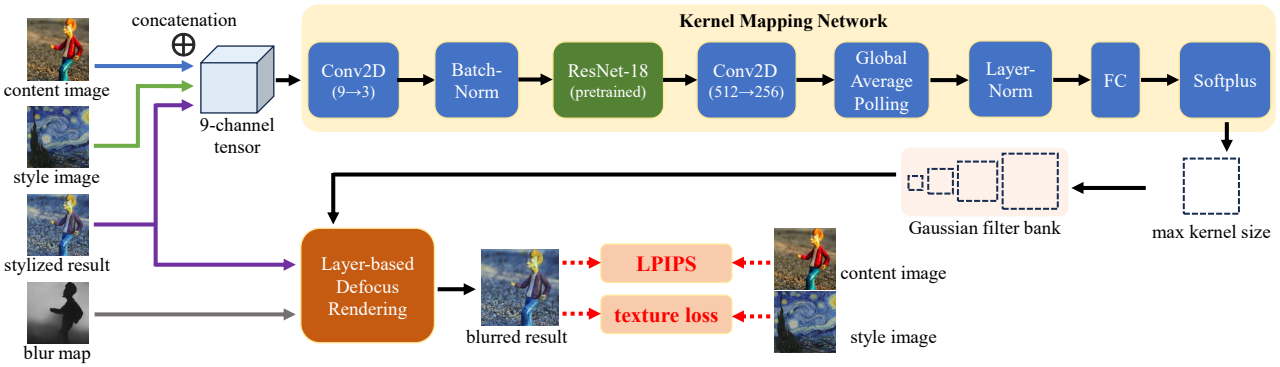

Figure 3: **An illustration of the architecture and optimization scheme of the Kernel Mapping Network.** The network takes as input a tensor formed by concatenating the content image, the style image, and the stylized output generated by a style transfer method. It predicts the maximum filter size for a Gaussian filter bank, which is then used to produce a blurred result through layer-wise Gaussian blurring. This blurred image is then compared to the content image using the LPIPS perceptual loss and to the style image using a texture loss. The total loss, combining both perceptual and texture losses, is used to optimize the predicted maximum kernel size.

imum kernel size, which is used to construct a Gaussian filter bank. This filter bank is then applied within a layer-based defocus rendering framework to simulate DoF effects, producing a blurred, stylized image. A loss function is subsequently used to evaluate the quality of the blurred result and guide the optimization of the network. To ensure that the synthesized defocus characteristics closely match those of the original content image, we employ the perceptual similarity metric, LPIPS [ZIE*18] as our loss function, measuring the visual distance between the blurred result and the content image. Compared to traditional metrics such as PSNR and SSIM, LPIPS better reflects human visual perception [BMT*18, BHMS19]. It evaluates perceptual differences between two images using a deep neural network (e.g., VGG [SZ15]), making it particularly sensitive to structural features like object shapes, textures, and edges. The LPIPS loss $\mathcal{L}_{\text{lpips}}(x,x_0)$ between the generated blurred result x and the original content image x_0 is defined as follows:

$$\mathcal{L}_{\text{lpips}}(x,x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \left\| w_l \odot \left(\hat{y}^l_{(hw)} - \hat{y}^l_{0(hw)}\right)\right\|_2^2, \quad (1)$$

where $\hat{y}^l_{(hw)}$ and $\hat{y}^l_{0(hw)}$ represent the unit-normalized feature vectors of dimension C at spatial location (h,w) in the l-th layer for images x and x_0, respectively. H_l and W_l denote the height and width of the feature map at layer l. The vector $w_l \in \mathbb{R}^C$ contains learned, channel-wise weights for layer l. The symbol $\odot$ represents element-wise (Hadamard) multiplication, and $\|\cdot\|_2$ denotes the standard Euclidean norm in $\mathbb{R}^C$.

When the loss function relies solely on LPIPS, it aligns the maximum blur kernel closely with the defocus properties of the content image. However, aesthetic preferences can vary, and some users may perceive the resulting stylized images as overly blurred. To address this, we incorporate an additional texture loss $\mathcal{L}_{\text{style}}$, introduced by Gatys *et al.* [GEB16]. This loss $\mathcal{L}_{\text{style}}(x,x_s)$ measures the stylistic discrepancy between the generated blurred output x and the reference style image x_s using CNN features. Specifically, it is defined as the sum of the squared Frobenius norms of the differences between their Gram matrices across all convolutional layers:

$$\mathcal{L}_{\text{style}}(x,x_s) = \sum_l \frac{w_l}{4N_l^2 M_l^2} \sum_{i,j} (G^l_{(ij)} - G^l_{s(ij)})^2. \quad (2)$$

Here, N_l and M_l represent the number of feature maps and their spatial dimensions in layer l, while w_l specifies the relative contribution of that layer to the overall loss. G^l and G^l_s denote the Gram matrices of images x and x_s, respectively. Each Gram matrix entry is defined as:

$$G^l_{(ij)} = \sum_k F^l_{ik} F^l_{jk}, \quad (3)$$

where F^l_{ik} and F^l_{jk} are the activations of the i-th and j-th feature maps at spatial location k in layer l, capturing the correlation between them. The weights w_l are assigned according to the original implementation, and feature representations are extracted from five convolutional layers of the VGG network [SZ15].

The total loss function is ultimately defined as a weighted combination of the LPIPS perceptual losswith the content image ($\mathcal{L}_{\text{lpips}}$, Eq. 1) and the texture loss with the style reference image ($\mathcal{L}_{\text{style}}$, Eq. 2), where a weighting factor λ controls their relative contributions:

$$\mathcal{L}_{total} = \mathcal{L}_{\text{lpips}} + \lambda \cdot \mathcal{L}_{\text{style}}. \quad (4)$$

Intuitively, setting $\lambda = 0$ prioritizes the preservation of the content image's defocus characteristics in the output. Conversely, based on our experimental observations, a value of $\lambda = 1000$ produces results that roughly match the original stylized image. Intermediate values of λ within the range $[0, 1000]$ offer a smooth trade-off between these two effects. In our default configuration, we set $\lambda = 0$, using only the perceptual loss, as it most effectively replicates the defocus properties of the original content image.

For the remaining network training hyperparameters, we use the Adam optimizer with a learning rate of 1×10^{-5}, optimizing the model for 100 epochs.

3.4. Layered-based defocus rendering

Our method builds upon the layer-based defocus rendering framework introduced by SteReFo [BHMS19], which simulates DoF effects on all-in-focus images. The key distinction is that their method first estimates a disparity map and then applies DoF based on a predefined focus plane. In contrast, our method leverages recent advances in learning defocus maps directly from images, eliminating the need for user-specified parameters.

Given a stylized image I produced by a style transfer method, along with the estimated blur map B and the optimal maximum filter size K predicted from the kernel mapping network, we compute the minimum blur magnitude $b_{\min}$, the maximum blur magnitude $b_{\max}$, and the range of blur value $b_{\max} - b_{\min}$. For layer-based defocus simulation, the blur map is discretized into $K+1$ distinct blur levels, where each level corresponds to a blur magnitude interval of $t = (b_{\max} - b_{\min})/(K+1)$. Each pixel in the blur map is assigned to a specific blur level based on its blur magnitude. Every blur level corresponds to a Gaussian filter of a particular size, ranging from 0 to K, which reflects the degree of blur. Together, these $K+1$ filters constitute the Gaussian filter bank.

Our method adopts a mask-based filtering strategy to combine individual layers. For each blur level $l \in [0, K]$, we construct a level mask M^l by identifying all pixels whose blur magnitudes lie within the interval $[b_{min} + l \cdot t, b_{min} + (l+1) \cdot t]$. Using this mask, the corresponding partial stylized image I^l is obtained through Hadamard multiplication: $I^l = M^l \odot I$. We also compute the average blur magnitude b^l for that level. Notably, the average blur magnitudes at the lowest (minimum blur) and highest (maximum blur) levels are denoted as $b^{\min}$ and $b^{\max}$, respectively.

To simulate defocus blur, we process blur levels sequentially, starting from the lowest and progressing to the highest. We maintain two accumulation buffers, M_s and I_s, for the progressively blurred result. For each level l, the Gaussian filter radius is determined as:

$$K \cdot (b^l - b^{\min})/(b^{\max} - b^{\min}). \tag{5}$$

This ensures that the lowest blur level maps to a zero-radius filter and the highest to the maximum radius K. A Gaussian filter of this size is applied to both the level mask M^l and the partial stylized image I^l, producing blurred outputs M^l_b and I^l_b. These are then accumulated into M_s and I_s, but only at pixels where the mask is non-zero. This strategy prevents out-of-focus regions from bleeding into in-focus areas, thereby effectively handling blur clipping. After processing all levels, the stylized image with defocus blur, I_b, is obtained by normalizing I_s with M_s. Full step-by-step implementation details are available in the supplementary materials.

4. Experiments

We implemented our method using PyTorch 2.4.0 with CUDA 11.8 on a PC with an Intel Core i5-12600K 3.70 GHz processor, 64GB RAM, and an NVIDIA GeForce RTX 4070 GPU. All results presented in this paper are at a resolution of 512×512 pixels. Estimating the blur map with the DMENet model [LLCL19] takes approximately 25 seconds. The runtime for producing the stylized output depends on the selected style transfer approach. For our defocus restoration, the process requires 30.5 seconds per image in addition to blur map estimation and stylization when using LPIPS alone as the loss function; incorporating texture loss increases the inference time to 36.6 seconds per image.

4.1. Integration with existing style transfer methods

We integrated our method with several existing style transfer techniques, including attention-aware [ZXW*24], depth-aware [IM22], aesthetic-aware [HJL*20], and parameterized brushstrokes-based approaches [KWHO21]. For each technique, we conducted a comparative analysis of the outputs with and without our enhancement. Perceptual differences between the content and stylized images were quantitatively assessed using the LPIPS metric (Eq. 1), where lower scores indicate higher perceptual similarity. Traditional metrics such as PSNR and SSIM do not adequately reflect the perceptual quality of defocused images [BMT*18, BHMS19]. Our evaluation was conducted on 80 content–style image pairs, with two representative examples for each approach shown in Figures 4 and 5. Additional results are provided in the supplementary material.

In Figures 4 and 5, the first column displays the style images alongside the estimated blur maps derived from the content image (second column). The third and fourth columns present the stylized outputs from the baseline method and our enhanced version, respectively. The fifth and sixth columns visualize the LPIPS error for the baseline and our stylized results, where blue represents lower perceptual error and red indicates higher error. For our method, we set the weight λ in Eq. 4 to zero, which yields results that most closely replicate the defocus effects of the content image.

Figure 4 presents the integration with the aesthetic-aware style transfer method proposed by Hu *et al.* [HJL*20], which separates aesthetic attributes into colour and texture components, enabling independent transfer from different reference images. Although Hu *et al.*'s method preserves some defocus blur in the stylized outputs, the blur often appears unnatural, where noticeable artifacts are present. In comparison, our method produces more realistic defocus effects that closely resemble the blur characteristics of the original content image.

The top part of Figure 5 presents the integration with the brushstroke-based style transfer method proposed by Kotovenko *et al.* [KWHO21]. On its own, this method does not preserve the defocus effects of the original content image, resulting in high perceptual error. Additionally, the distinction between focused and defocused regions becomes less clear after style transfer. By incorporating our approach, defocus characteristics are effectively reintroduced into the stylized image, significantly enhancing perceptual consistency with the content image and restoring a clear separation between focused and defocused areas.

The middle part of Figure 5 presents the integration with the depth-aware style transfer method introduced by Ioannou and Maddock [IM22]. Their technique successfully preserves the structural consistency in the stylized results using depth estimation. However, when the content image contains a highly defocused background, their method can result in an overly detailed and textured backdrop. By introducing defocus effects explicitly through our approach, the

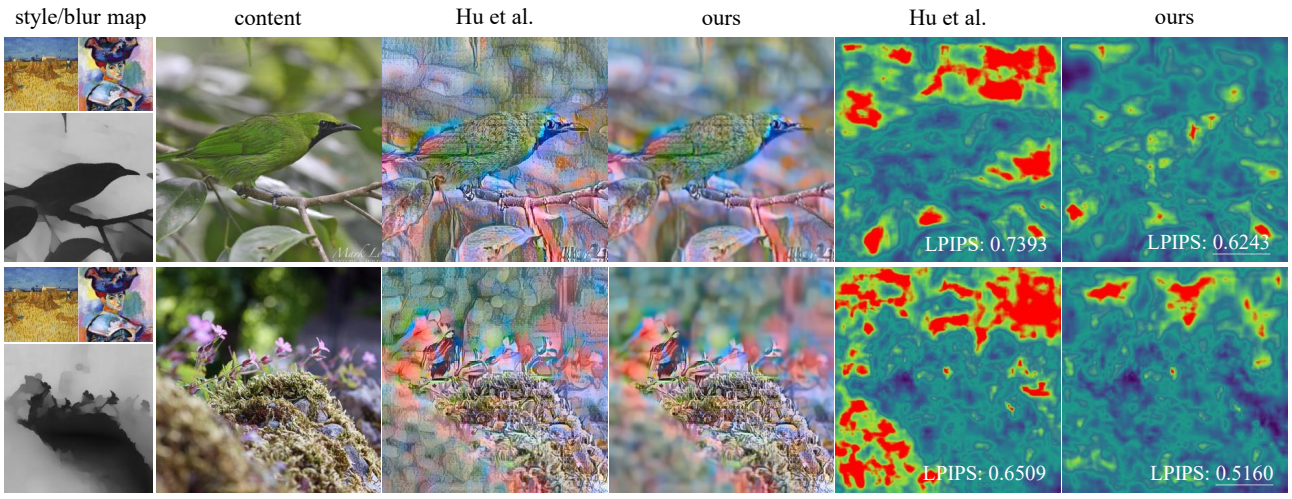

Figure 4: **Integration with the aesthetic-aware style transfer method [HJL*20].** In the first column, the texture (top left) and color (top right) style reference images are displayed, along with the blur map (bottom) estimated from the content image shown in the second column. While the method by Hu *et al*.'s retains some defocus blur, the results often exhibit unnatural artifacts. In contrast, our method generates defocus effects that more closely align with those in the original content images. The LPIPS scores and visualizations in the last two columns further demonstrate that our results are perceptually more similar to the original content images.

background becomes appears more subdued, effectively enhancing focus on the main subjects.

Lastly, we present the integration with S2WAT [ZXW*24], an attention-based style transfer method that employs a hierarchical vision transformer, in the bottom part of Figure 5. In the outputs generated by S2WAT, the stylized background exhibit artifacts (top example) and excessively detailed textures (bottom example), which compromise the intended defocused aesthetic of the original content image. Our method mitigates these issues by suppressing high-frequency details in the background, resulting in a visual appearance more consistent with the original image.

Our results markedly enhance the aesthetic quality of the stylized images. In the content examples, sharp foreground objects are contrasted with softly blurred backgrounds: a composition shown to enhance visual appeal in prior studies [KTJ06, KSL*16]. This separation is diminished in the initial stylized results, which often appear cluttered and lack visual focus. At the same time, as noted by Datta *et al*. [DJLW06], excessive blur can also reduce aesthetic quality, underscoring the importance of balance. Our method achieves this by estimating an optimal filter size, accurately reproducing the original blur characteristics of the content image and restoring the prominence of the foreground.

In each example, the pixel-wise LPIPS error visualizations (fifth and sixth columns) clearly demonstrate that our approach significantly reduces perceptual error, particularly in out-of-focus areas, leading to a notable improvement in overall perceptual fidelity. Across 80 test cases, the average LPIPS error of the original stylized outputs is 0.6567, whereas our method achieves 0.5338.

4.2. Comparison with defocus adaptive style transfer

Figure 6 compares our results with those of Cao *et al*. [CCJT23], who utilize a stroke pyramid approach to emphasize defocus effects in stylized images. Since neither source code nor pretrained models are publicly available, we rely on the visual results from their publication for comparison. Their outputs, shown in the third column, demonstrate how varying stroke sizes can enhance semantically significant regions. In contrast, when their outputs are combined with our method, as displayed in the fourth column, a clearer distinction emerges between focused and defocused areas. Additionally, the LPIPS error decreases, indicating a closer perceptual match to the original content image.

4.3. Trade-offs between content consistency and style fidelity

Figure 7 compares results generated with different values of the weight λ in Equation 4, which mediates the trade-off between content consistency and texture fidelity to accommodate various aesthetic preferences. For this experiment, we integrated our method into S2WAT [ZXW*24], an attention-based style transfer technique. As the style loss weight increases, more background details is preserved, demonstrating that users can adjust λ to balance depth-of-field blur against style emphasis. When $\lambda = 0$, the LPIPS error relative to the content image is minimized, since the loss prioritizes perceptual similarity to the original content.

4.4. User studies

While LPIPS aligns more closely with human visual perception than other metrics, it does not fully reflect perceived quality. To complement this, we conducted a user study with 54 participants. Each participant evaluated 16 image sets, comprising four randomly chosen sets for each integrated method:

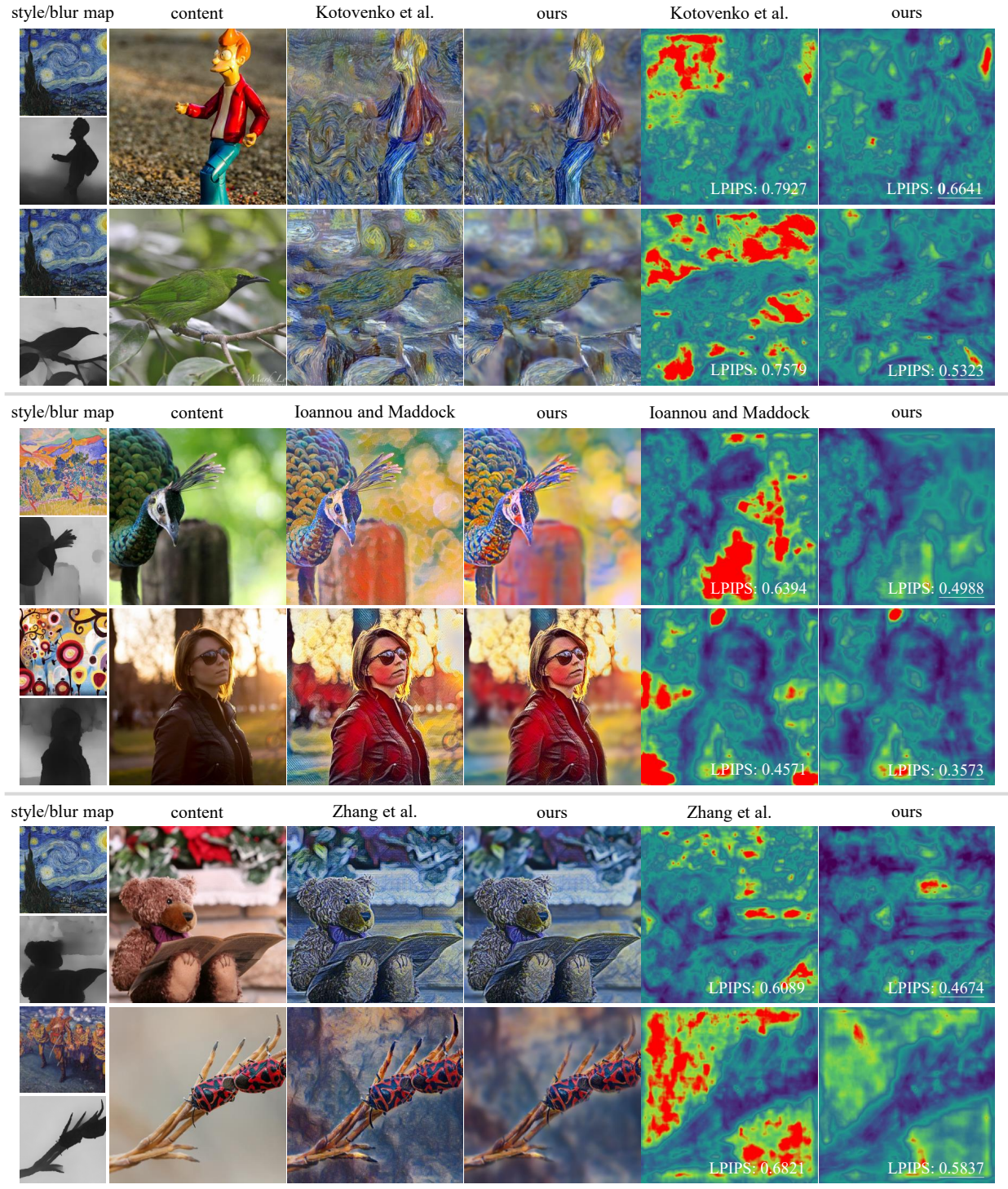

Figure 5: **Integration with the brushstroke-based [KWHO21] (top), depth-aware [IM22] (middle), and attention-aware [ZXW*24] (bottom) style transfer methods.** The first column displays the style image (top) alongside the blur map (bottom), estimated from the content image shown in the second column. The third column presents the outputs from the original stylized approaches, which often weaken the intended focus by reducing the separation between sharp and blurred regions and, in some cases, introducing artifacts or unnaturally sharp backgrounds. Our enhanced results, shown in the fourth column, address these issues by suppressing distractions in defocused regions and better preserving the original emphasis. The final two columns report LPIPS scores and provide a side-by-side error comparison between the baseline stylized outputs and our enhanced results.

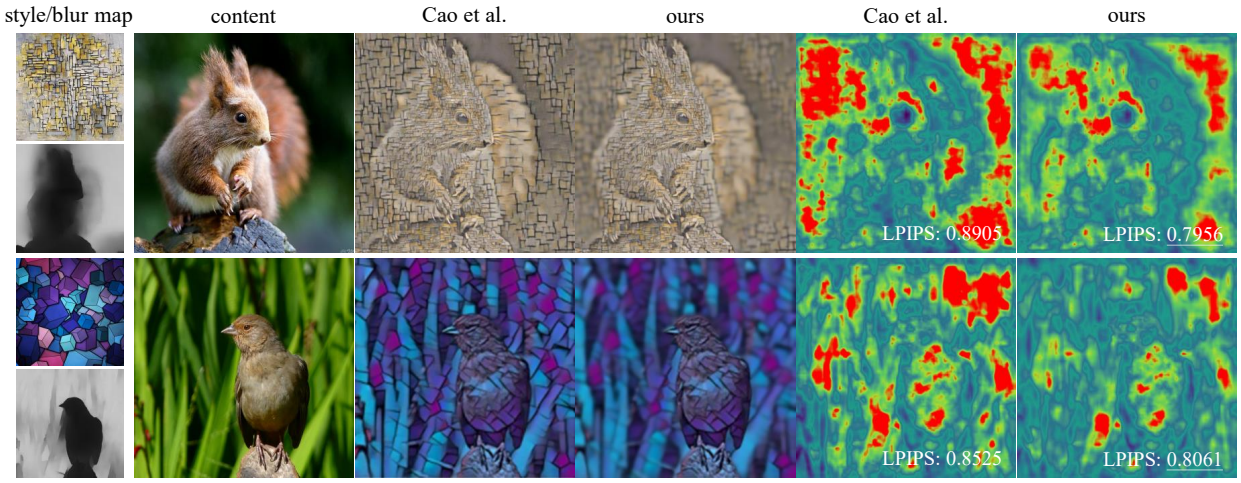

Figure 6: **Comparisons with Cao *et al.*'s defocus-aware method [CCJT23].** The first column shows the style image (top) and the corresponding blur map (bottom), which is estimated from the content image shown in the second column. Unlike the method by Cao *et al.*, which adjusts stroke sizes according to the level of defocus, our approach directly preserves the defocus characteristics of the original content image by applying background blurring.

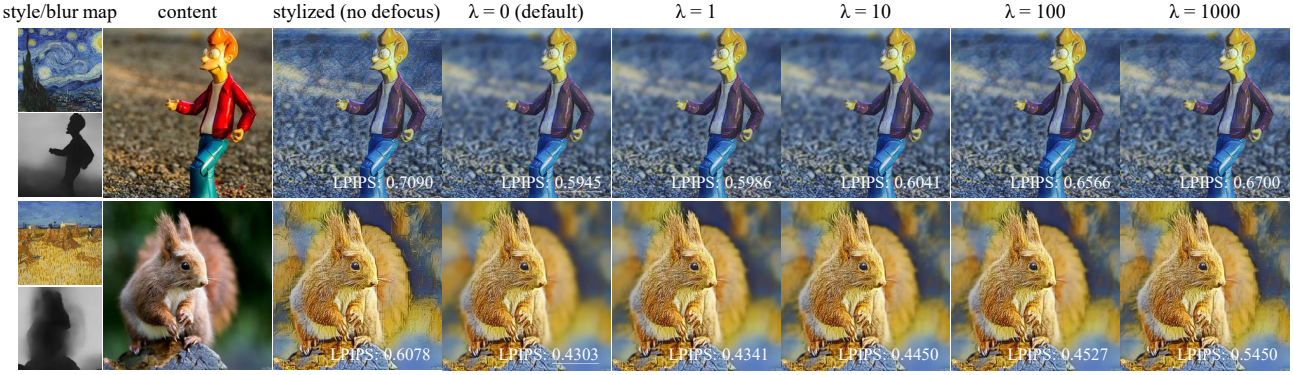

Figure 7: Comparisons of results obtained using different weights to balance perceptual loss and texture loss. The first column shows the style image (top) and the blur map (bottom) estimated from the content image. The base stylized image is generated by S2WAT [ZXW*24], which our method builds upon to further enhance results by reintroducing defocus effects. A weight of zero prioritizes minimizing LPIPS perceptual error, whereas increasing the weight gradually enhances texture details in the output.

aesthetic-aware [HJL*20], brushstroke-based [KWHO21], depth-aware [IM22], and attention-aware [ZXW*24] style transfer methods. For each set, participants were shown the original content image, the style image(s), the original stylized output, and our enhanced result with reinstated defocus. They were then asked to choose between the original and enhanced stylized images in response to two questions: (1) Which image better preserved the structure of the content image? (2) Which image is more visually appealing?

Figure 8 shows the results of the user study. Out of 864 responses (54 participants $\times$ 16 sets), 88.6% agreed that our method better preserved the structural details of the content image, attributable to the restored defocus effects. Nearly 70% of participants also rated our results as more visually appealing. Overall, across all integrated methods, our approach consistently received higher preference scores than the original stylized outputs.

It is important to note that the evaluation was conducted using the default configuration without style loss (*i.e.*, $\lambda = 0$ in Eq. 4). Users seeking stronger stylization with less blurring can adjust this weight to achieve their desired balance, as discussed in Sec. 4.3.

4.5. Ablation study

We conducted an ablation study to evaluate the impact of removing specific components from our network, leading to two simplified variants: one without the pretrained ResNet-18 backbone and another without the softplus activation. The comparison is presented in Figure 9. ResNet-18 contributes rich semantic features while keeping the parameter count relatively low, thereby improving results with minimal overhead. In its absence, the outputs often exhibit overly sharp textures in defocused regions. Removing softplus, meanwhile, can introduce artifacts due to negative kernel

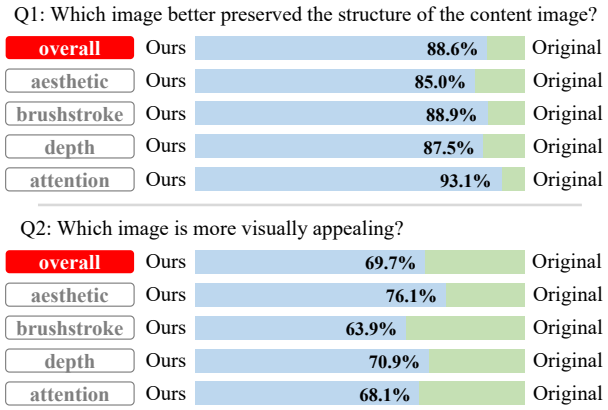

Figure 8: **User study.** The majority of participants agreed that our method successfully preserves the content structure, while nearly 70 preferred our results for visual appeal. These findings were consistent across all integrated methods.

sizes from the fully connected layer. Across all 80 test cases, the average LPIPS scores for the full model, the ResNet-18–free variant, and the softplus-free variant are 0.5338, 0.5739, and 0.6462, respectively. These results confirm that the full model achieves the best performance in terms of both visual quality and LPIPS scores.

To further validate the effectiveness of our method, we compare our results against those obtained with different kernel sizes. Because visually determining the optimal filter size is impractical, we vary the maximum kernel size predicted by our kernel mapping network for evaluation. As shown in Figure 10, the kernel size K selected by our method produces defocus effects that most closely resemble the original content image. Smaller kernels ($K/4$ and $K/2$) result in overly sharp textures in defocused regions, while larger kernels ($2K$ and $4K$) yield overly smooth outputs that lack distinct bokeh characteristics. Moreover, our approach achieves the best LPIPS scores across all configurations, confirming that the kernel mapping network selects an optimal filter size.

5. Conclusion

Defocused effects are common found in photographs and images, as a natural result of camera optics or as a deliberate technique to highlight specific subjects. However, most existing style transfer approaches tend to overlook these effects, resulting in stylized outputs that lack the original defocused appearance. To address this limitation, we propose a novel post-processing method that can be seamlessly integrated with any existing style transfer technique by reintroducing defocus to the stylized image. Our approach involves estimating a blur map from the content image, predicts the optimal maximum kernel size using a neural network to construct a Gaussian filter bank, and employs layer-based defocus rendering to simulate realistic defocus blur. To support varying aesthetic preferences, we design a loss function that combines perceptual similarity to the content image and texture fidelity to the style image. Extensive experiments demonstrate that our method significantly improves the quality of stylized images, especially when the original content image includes defocus effects.

One limitation of our method is its strong reliance on accurate blur estimation. Consequently, an imprecise blur map can result in suboptimal outputs. An example is shown in Figure 11, where the blur map incorrectly estimates the blur on the sleeve, which negatively impacts the final output. A potential direction for future work is to design a more robust and reliable blur estimation method, potentially by integrating the blur estimation directly into the defocus simulation pipeline.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was supported in part by the National Science and Technology Council (NSTC) under grants 113-2221-E-305-010-MY3.

References

[AAB18] ATAPOUR-ABARGHOUEI A., BRECKON T. P.: Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 2800–2810. doi:10.1109/CVPR.2018.00296. 3

[BD07] BAE S., DURAND F.: Defocus magnification. *Computer Graphics Forum 26*, 3 (2007), 571–579. doi:https://doi.org/10.1111/j.1467-8659.2007.01080.x. 3

[BHMS19] BUSAM B., HOG M., MCDONAGH S., SLABAUGH G.: SteReFo: Efficient image refocusing with stereo vision. In *IEEE/CVF International Conference on Computer Vision (ICCV) Workshops* (2019), pp. 3295–3304. doi:10.1109/ICCVW.2019.00411. 3, 4, 5, 6

[BMT*18] BLAU Y., MECHREZ R., TIMOFTE R., MICHAELI T., ZELNIK-MANOR L.: The 2018 pirm challenge on perceptual image super-resolution. In *European Conference on Computer Vision (ECCV) Workshops* (September 2018). 5, 6

[CCJT23] CAO J., CHEN Z., JIN M., TIAN Y.: An improved defocusing adaptive style transfer method based on a stroke pyramid. *PLOS ONE 18*, 4 (04 2023), 1–19. doi:10.1371/journal.pone.0284742. 2, 3, 7, 9

[CJW*21] CHENG J., JAISWAL A., WU Y., NATARAJAN P., NATARAJAN P.: Style-aware normalized loss for improving arbitrary style transfer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 134–143. doi:10.1109/CVPR46437.2021.00020. 3

[CLW*20] CHENG M.-M., LIU X.-C., WANG J., LU S.-P., LAI Y.-K., ROSIN P. L.: Structure-preserving neural style transfer. *IEEE Transactions on Image Processing 29* (2020), 909–920. doi:10.1109/TIP.2019.2936746. 2

[CP20] CUN X., PUN C.-M.: Defocus blur detection via depth distillation. In *Proc. European Conference on Computer Vision (ECCV)* (2020), p. 747–763. doi:10.1007/978-3-030-58601-0_44. 3

[CS16] CHEN T. Q., SCHMIDT M.: Fast patch-based style transfer of arbitrary style, 2016. URL: https://arxiv.org/abs/1612.04337, arXiv:1612.04337. 2

[CWC20] CHANG H.-Y., WANG Z., CHUANG Y.-Y.: Domain-specific mappings for generative adversarial style transfer. In *Proc. European Conference on Computer Vision (ECCV)* (2020), pp. 573–589. 3

[CWJ*23] CHENG J., WU Y., JAISWAL A., ZHANG X., NATARAJAN P., NATARAJAN P.: User-controllable arbitrary style transfer via entropy regularization. In *Proc. International Conference on Neural Information Processing Systems (AAAI)* (2023). doi:10.1609/aaai.v37i1.25117. 3

[CZLY23] CHEN H., ZHAO L., LI J., YANG J.: TSSAT: Two-stage statistics-aware transformation for artistic style transfer. In *Proc. ACM*

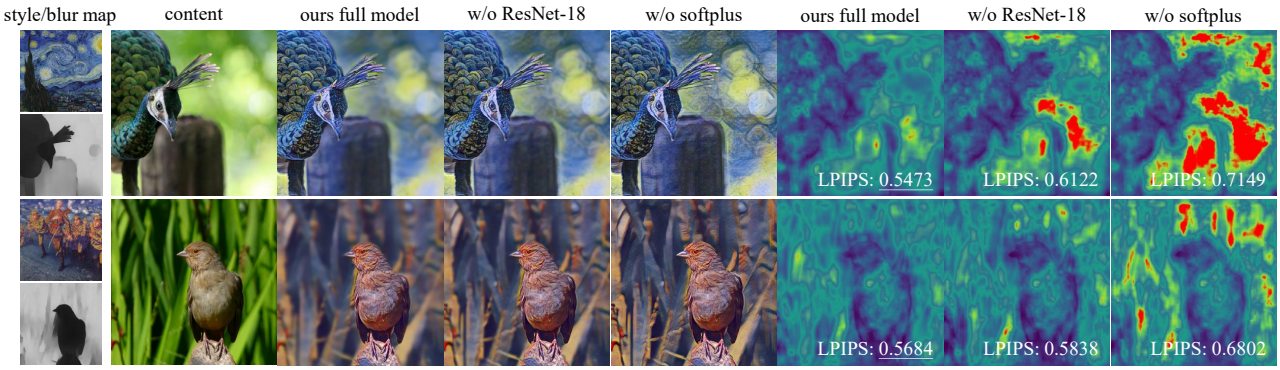

Figure 9: **Ablation studies.** Omitting ResNet-18 or softplus function results in suboptimal results, such as overly sharp textures in defocused regions and instability. The full model delivers the best performance.

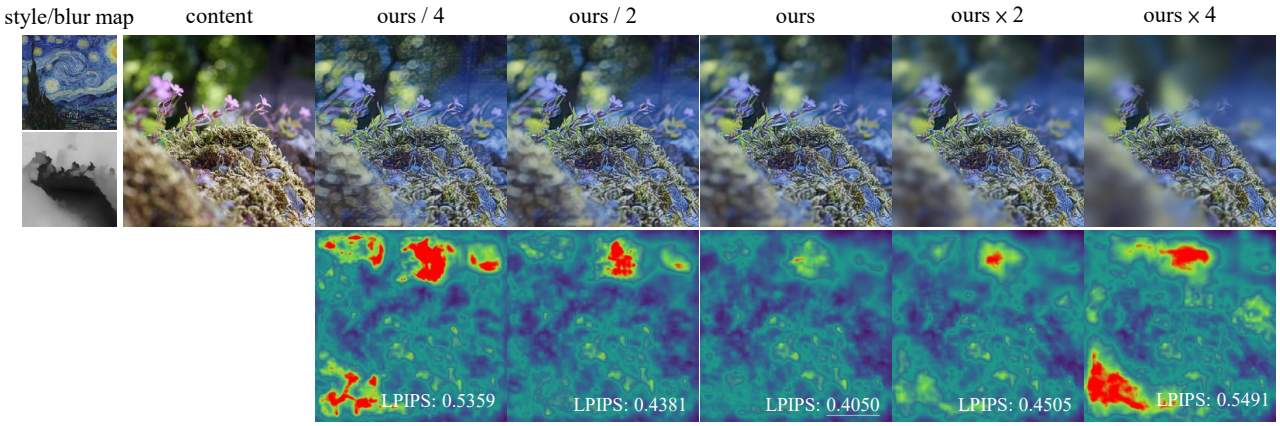

Figure 10: Comparison of our result against those produced using different maximum kernel sizes for building the Gaussian filter bank. The kernel size predicted by our kernel mapping network most accurately replicates the defocus present in the content image, avoiding both overly sharp textures and excessive blurring. Furthermore, our method achieves the lowest LPIPS error among all tested configurations.

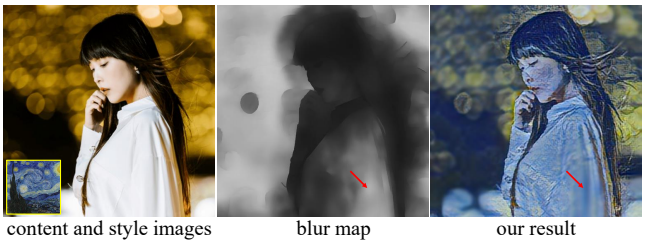

Figure 11: Our method could generate inaccurate defocus effects when the blur estimation is incorrect, as seen in the overblurring of the sleeve in the blur map.

International Conference on Multimedia (2023), p. 6878–6887. doi:10.1145/3581783.3611819. 2, 3

[CZW*21a] CHEN H., ZHAO L., WANG Z., ZHANG H., ZUO Z., LI A., XING W., LU D.: Artistic style transfer with internal-external learning and contrastive learning. In *Proc. International Conference on Neural Information Processing Systems (AAAI)* (2021). 2

[CZW*21b] CHEN H., ZHAO L., WANG Z., ZHANG H., ZUO Z., LI A., XING W., LU D.: DualAST: Dual style-learning networks for artistic style transfer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 872–881. doi:10.1109/CVPR46437.2021.00093. 2

[CZZ*21] CHEN H., ZHAO L., ZHANG H., WANG Z., ZUO Z., LI A., XING W., LU D.: Diverse image style transfer via invertible cross-space mapping. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 14860–14869. doi:10.1109/ICCV48922.2021.01461. 3

[DDS*09] DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L.: ImageNet: A large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2009), pp. 248–255. doi:10.1109/CVPR.2009.5206848. 4

[DJLW06] DATTA R., JOSHI D., LI J., WANG J. Z.: Studying aesthetics in photographic images using a computational approach. In *Proc. European Conference on Computer Vision (ECCV)* (2006), pp. 288–301. 7

[DTD*20] DENG Y., TANG F., DONG W., SUN W., HUANG F., XU C.: Arbitrary style transfer via multi-adaptation network. In *Proc. ACM International Conference on Multimedia* (2020), p. 2719–2727. doi:10.1145/3394171.3414015. 2

[DTD*21] DENG Y., TANG F., DONG W., HUANG H., MA C., XU C.: Arbitrary video style transfer via multi-channel correlation. In *Proc. International Conference on Neural Information Processing Sys-*

tems (AAAI) (May 2021), pp. 1210–1217. doi:10.1609/aaai.v35i2.16208. 2

[DTD*22] DENG Y., TANG F., DONG W., MA C., PAN X., WANG L., XU C.: StyTr2: Image style transfer with transformers. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 11316–11326. doi:10.1109/CVPR52688.2022.01104. 2

[GEB16] GATYS L. A., ECKER A. S., BETHGE M.: Image style transfer using convolutional neural networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 2414–2423. doi:10.1109/CVPR.2016.265. 2, 5

[HB17] HUANG X., BELONGIE S.: Arbitrary style transfer in real-time with adaptive instance normalization. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2017), pp. 1510–1519. doi:10.1109/ICCV.2017.167. 2

[HJL*20] HU Z., JIA J., LIU B., BU Y., FU J.: Aesthetic-aware image style transfer. In *Proc. ACM International Conference on Multimedia* (2020), p. 3320–3329. doi:10.1145/3394171.3413853. 1, 2, 3, 6, 7, 9

[HJL*23] HONG K., JEON S., LEE J., AHN N., KIM K., LEE P., KIM D., UH Y., BYUN H.: AesPA-Net: Aesthetic pattern-aware style transfer networks. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 22701–22710. doi:10.1109/ICCV51070.2023.02080. 3

[HJY*21] HONG K., JEON S., YANG H., FU J., BYUN H.: Domain-aware universal style transfer. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 14589–14597. doi:10.1109/ICCV48922.2021.01434. 3

[IM22] IOANNOU E., MADDOCK S.: Depth-aware neural style transfer using instance normalization. In *Proc. Computer Graphics and Visual Computing* (2022). doi:10.2312/cgvc.20221165. 1, 2, 6, 8, 9

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *Proc. European Conference on Computer Vision (ECCV)* (2016), pp. 694–711. 2

[JLY*18] JING Y., LIU Y., YANG Y., FENG Z., YU Y., TAO D., SONG M.: Stroke controllable fast style transfer with adaptive receptive fields. In *Proc. European Conference on Computer Vision (ECCV)* (September 2018). 3

[JLYP13] JIANG P., LING H., YU J., PENG J.: Salient region detection by ufo: Uniqueness, focusness and objectness. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2013), pp. 1976–1983. doi:10.1109/ICCV.2013.248. 3

[KJ18] KARAALI A., JUNG C. R.: Edge-based defocus blur estimation with adaptive scale selection. *IEEE Transactions on Image Processing 27*, 3 (2018), 1126–1137. doi:10.1109/TIP.2017.2771563. 3

[KS07] KRAUS M., STRENGERT M.: Depth-of-field rendering by pyramidal image processing. *Computer Graphics Forum 26*, 3 (2007), 645–654. doi:10.1111/j.1467-8659.2007.01088.x. 4

[KSL*16] KONG S., SHEN X., LIN Z., MECH R., FOWLKES C.: Photo aesthetics ranking network with attributes and content adaptation. In *Proc. European Conference on Computer Vision (ECCV)* (2016), pp. 662–679. 2, 7

[KSP*18] KIM B., SON H., PARK S.-J., CHO S., LEE S.: Defocus and motion blur detection with deep contextual features. *Computer Graphics Forum* (2018). doi:10.1111/cgf.13567. 3

[KTJ06] KE Y., TANG X., JING F.: The design of high-level features for photo quality assessment. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2006), pp. 419–426. doi:10.1109/CVPR.2006.303. 2, 7

[KWHO21] KOTOVENKO D., WRIGHT M., HEIMBRECHT A., OMMER B.: Rethinking style transfer: From pixels to parameterized brushstrokes. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 12191–12200. doi:10.1109/CVPR46437.2021.01202. 1, 2, 3, 6, 8, 9

[LCLR17] LIU X.-C., CHENG M.-M., LAI Y.-K., ROSIN P. L.: Depth-aware neural style transfer. In *Proc. Symposium on Non-Photorealistic Animation and Rendering* (2017). doi:10.1145/3092919.3092924. 2

[LFY*17a] LI Y., FANG C., YANG J., WANG Z., LU X., YANG M.-H.: Diversified texture synthesis with feed-forward networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 266–274. doi:10.1109/CVPR.2017.36. 2

[LFY*17b] LI Y., FANG C., YANG J., WANG Z., LU X., YANG M.-H.: Universal style transfer via feature transforms. In *Proc. International Conference on Neural Information Processing Systems (NeurIPS)* (2017), p. 385–395. 2

[LH22] LIAO Y.-S., HUANG C.-R.: Semantic context-aware image style transfer. *IEEE Transactions on Image Processing 31* (2022), 1911–1923. doi:10.1109/TIP.2022.3149237. 2

[LLCL19] LEE J., LEE S., CHO S., LEE S.: Deep defocus map estimation using domain adaptation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 12214–12222. doi:10.1109/CVPR.2019.01250. 2, 3, 4, 6

[LLW*23] LI D., LUO H., WANG P., WANG Z., LIU S., WANG F.: Frequency domain disentanglement for arbitrary neural style transfer. In *Proc. International Conference on Neural Information Processing Systems (AAAI)* (2023). doi:10.1609/aaai.v37i1.25212. 2

[LTD*21] LIN M., TANG F., DONG W., LI X., XU C., MA C.: Distribution aligned multimodal and multi-domain image stylization. *ACM Transactions Multimedia Computing, Communications, and Applications 17*, 3 (July 2021). doi:10.1145/3450525. 3

[LWH24] LIU X.-C., WU Y.-C., HALL P.: Painterly style transfer with learned brush strokes. *IEEE Transactions on Visualization and Computer Graphics 30*, 9 (2024), 6309–6320. doi:10.1109/TVCG.2023.3332950. 3

[LZ22] LIU S., ZHU T.: Structure-guided arbitrary style transfer for artistic image and video. *IEEE Transactions on Multimedia 24* (2022), 1299–1312. doi:10.1109/TMM.2021.3063605. 2

[PCL*22] PENG J., CAO Z., LUO X., LU H., XIAN K., ZHANG J.: BokehMe: When neural rendering meets classical rendering. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 16262–16271. doi:10.1109/CVPR52688.2022.01580. 3

[PCL*25] PENG J., CAO Z., LUO X., XIAN K., TANG W., ZHANG J., LIN G.: BokehMe++: Harmonious fusion of classical and neural rendering for versatile bokeh creation. *IEEE Trans. on Pattern Analysis and Machine Intelligence 47*, 3 (2025), 1530–1547. doi:10.1109/TPAMI.2024.3501739. 3

[PL19] PARK D. Y., LEE K. H.: Arbitrary style transfer with style-attentional networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2019), pp. 5873–5881. doi:10.1109/CVPR.2019.00603. 2

[RCLL22] RUAN L., CHEN B., LI J., LAM M.: Learning to deblur using light field generated and real defocus images. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 16283–16292. doi:10.1109/CVPR52688.2022.01582. 3

[RMN*19] R. I. A. L., MENON L. T., N. M. C. O. P., KOERICH A. L., JR A. S. B.: Style transfer applied to face liveness detection with user-centered models, 2019. URL: https://arxiv.org/abs/1907.07270, arXiv:1907.07270. 3

[SBC12] SAAD M. A., BOVIK A. C., CHARRIER C.: Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE Transactions on Image Processing 21*, 8 (2012), 3339–3352. doi:10.1109/TIP.2012.2191563. 3

[SCK*23] SEIZINGER T., CONDE M. V., KOLMET M., BISHOP T. E., TIMOFTE R.: Efficient multi-lens bokeh effect rendering and transformation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2023). doi:10.1109/CVPRW59228.2023.00165. 3

[SGW*18] SRINIVASAN P. P., GARG R., WADHWA N., NG R., BARRON J. T.: Aperture supervision for monocular depth estimation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2018), pp. 6393–6401. doi:10.1109/CVPR.2018.00669. 3

[SWS*17] SRINIVASAN P. P., WANG T., SREELAL A., RAMAMOORTHI R., NG R.: Learning to synthesize a 4d rgbd light field from a single image. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2017), pp. 2262–2270. doi:10.1109/ICCV.2017.246. 3

[SXJ14] SHI J., XU L., JIA J.: Discriminative blur detection features. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 2965–2972. doi:10.1109/CVPR.2014.379. 3

[SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)* (2015). 5

[TLL*23] TANG H., LIU S., LIN T., HUANG S., LI F., HE D., WANG X.: Master: Meta style transformer for controllable zero-shot and few-shot artistic style transfer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), pp. 18329–18338. doi:10.1109/CVPR52729.2023.01758. 3

[TM98] TOMASI C., MANDUCHI R.: Bilateral filtering for gray and color images. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (1998), pp. 839–846. doi:10.1109/ICCV.1998.710815. 2, 4

[TWH*16] TANG C., WU J., HOU Y., WANG P., LI W.: A spectral and spatial approach of coarse-to-fine blurred image region detection. *IEEE Signal Processing Letters 23*, 11 (2016), 1652–1656. doi:10.1109/LSP.2016.2611608. 3

[ULVL16] ULYANOV D., LEBEDEV V., VEDALDI A., LEMPITSKY V.: Texture networks: feed-forward synthesis of textures and stylized images. In *Proc. International Conference on International Conference on Machine Learning (ICML)* (2016), p. 1349–1357. 2

[WGJ*18] WADHWA N., GARG R., JACOBS D. E., FELDMAN B. E., KANAZAWA N., CARROLL R., MOVSHOVITZ-ATTIAS Y., BARRON J. T., PRITCH Y., LEVOY M.: Synthetic depth-of-field with a single-camera mobile phone. *ACM Trans. Graph. 37*, 4 (July 2018). doi:10.1145/3197517.3201329. 3

[WHSX21] WU X., HU Z., SHENG L., XU D.: StyleFormer: Real-time arbitrary style transfer via parametric style composition. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), pp. 14598–14607. doi:10.1109/ICCV48922.2021.01435. 2

[WLZF22] WANG Q., LI S., ZHANG X., FENG G.: Multi-granularity brushstrokes network for universal style transfer. *ACM Transactions Multimedia Computing, Communications, and Applications 18*, 4 (Mar. 2022). doi:10.1145/3506710. 3

[WSZ*18] WANG L., SHEN X., ZHANG J., WANG O., LIN Z., HSIEH C.-Y., KONG S., LU H.: DeepLens: shallow depth of field from a single image. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 37*, 6 (Dec. 2018). doi:10.1145/3272127.3275013. 3

[WSZ*20] WU Z., SONG C., ZHOU Y., GONG M., HUANG H.: EFANet: Exchangeable feature alignment network for arbitrary style transfer. In *Proc. International Conference on Neural Information Processing Systems (AAAI)* (Apr. 2020), pp. 12305–12312. doi:10.1609/aaai.v34i07.6914. 2

[XQJ17] XU G., QUAN Y., JI H.: Estimating defocus blur via rank of local patches. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2017), pp. 5381–5389. doi:10.1109/ICCV.2017.574. 3

[XSS23] XU Z., SANGINETO E., SEBE N.: StylerDALLE: Language-guided style transfer using a vector-quantized tokenizer of a large-scale generative model. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 7567–7577. doi:10.1109/ICCV51070.2023.00699. 3

[YE16] YI X., ERAMIAN M.: LBP-based segmentation of defocus blur. *IEEE Transactions on Image Processing 25*, 4 (2016), 1626–1638. doi:10.1109/TIP.2016.2528042. 3

[YLY*16] YANG Y., LIN H., YU Z., PARIS S., YU J.: Virtual DSLR: High quality dynamic depth-of-field synthesis on mobile platforms. *Electronic Imaging 28*, 18 (2016), 1–9. doi:10.2352/ISSN.2470-1173.2016.18.DPMI-031. 3

[YRX*19] YAO Y., REN J., XIE X., LIU W., LIU Y.-J., WANG J.: Attention-aware multi-stroke style transfer. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 1467–1475. doi:10.1109/CVPR.2019.00156. 2

[YTy*11] YU Z., THORPE C., YU X., GRAUER-GRAY S., LI F., YU J.: Dynamic depth of field on live video streams: A stereo solution. In *Proc. Computer Graphics International* (01 2011). 3

[ZD18] ZHANG H., DANA K.: Multi-style generative network for real-time transfer. In *Proc. European Conference on Computer Vision Workshops* (September 2018). 2

[ZDCY23] ZHANG C., DAI Z., CAO P., YANG J.: Edge enhanced image style transfer via transformers. In *Proc. ACM International Conference on Multimedia Retrieval* (2023), p. 105–114. doi:10.1145/3591106.3592257. 2

[ZHW*23] ZHU M., HE X., WANG N., WANG X., GAO X.: All-to-key attention for arbitrary style transfer. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 23052–23062. doi:10.1109/ICCV51070.2023.02112. 2

[ZHW*24] ZHAO W., HU G., WEI F., WANG H., HE Y., LU H.: Attacking defocus detection with blur-aware transformation for defocus deblurring. *IEEE Transactions on Multimedia 26* (2024), 5450–5460. doi:10.1109/TMM.2023.3334023. 3

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), pp. 586–595. doi:10.48550/arXiv.1801.03924. 2, 5

[ZLZ23] ZHENG Z., LIU J., ZHENG N.: P^2-GAN: Efficient stroke style transfer using single style image. *IEEE Transactions on Multimedia 25* (2023), 6000–6012. doi:10.1109/TMM.2022.3203220. 3

[ZLZY22] ZHAO F., LU H., ZHAO W., YAO L.: Image-scale-symmetric cooperative network for defocus blur detection. *IEEE Transactions on Circuits and Systems for Video Technology 32*, 5 (2022), 2719–2731. doi:10.1109/TCSVT.2021.3095347. 3

[ZS11] ZHUO S., SIM T.: Defocus map estimation from a single image. *Pattern Recognition 44*, 9 (2011), 1852–1858. doi:https://doi.org/10.1016/j.patcog.2011.03.009. 3

[ZSL21] ZHAO W., SHANG C., LU H.: Self-generated defocus blur detection via dual adversarial discriminators. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 6929–6938. doi:10.1109/CVPR46437.2021.00686. 3

[ZXW*24] ZHANG C., XU X., WANG L., DAI Z., YANG J.: S2WAT: image style transfer via hierarchical vision transformer using strips window attention. In *Proc. International Conference on Neural Information Processing Systems (AAAI)* (2024). doi:10.1609/aaai.v38i7.28529. 1, 2, 6, 7, 8, 9

[ZZL*22] ZUO Z., ZHAO L., LIAN S., CHEN H., WANG Z., LI A., XING W., LU D.: Style fader generative adversarial networks for style degree controllable artistic style transfer. In *Proc. International Joint Conference on Artificial Intelligence* (7 2022), pp. 5002–5009. doi:10.24963/ijcai.2022/693. 3