TCDNet: Texture and Color Dynamic Network for Image Harmonization

Shan Yue*[1][0009−0000−7664−0440], Hai Huang* **[1][0000−0002−3637−1378], Zhenqi Tang[1][0009−0008−1179−9213], Yutong Zheng[1][0009−0006−7439−5385], and Zhou Fang[1][0009−0007−6271−4622]

School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
{yshan,huanghai,tangzq,zhengyutong,zhoufang1}@bupt.edu.cn

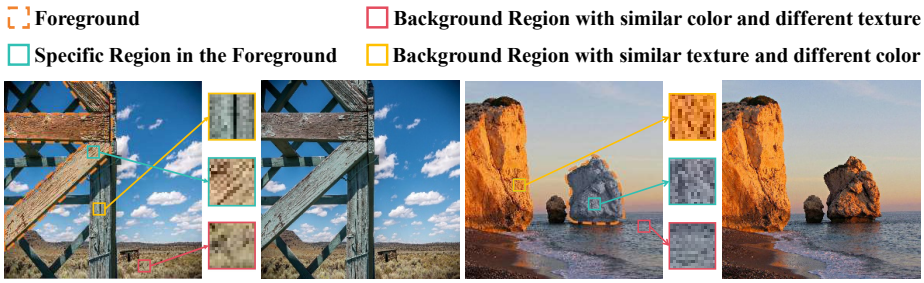
Abstract. Image harmonization plays an important role in computer vision, enhancing the realism of composite images. However, existing work focus on color adjustments while neglecting the impact of texture on color coherence. To address this issue, we propose the Texture and Color Dynamic Network (TCDNet), a new dual-encoder single-decoder architecture. Our TCDNet aims to achieve image harmonization through a unified texture-color perspective from both foreground and background regions. Specifically, we employ two task-specific encoders, *i.e.*, a texture encoder and a color encoder, to separately extract texture and color features. Subsequently, we designed a Texture based Color Transfer (TBCT) module to align the color representation of the foreground with that of the background, leveraging texture-based cues. Within TBCT, attention mechanisms and position encoding refine textural details, ensuring consistent texture alignment of foreground and background. During decoding, we propose a Color Dynamic (CoDy) module to dynamically adapt kernels to navigate and reinforce color correlations across varying input conditions. This synergistic interplay between texture and color dynamics enables TCDNet to navigate the complex landscape of image harmonization with high precision. We conducted extensive experiments on synthetic and real data to demonstrate the competitive performance of our method when compared to state-of-the-art (SOTA) supervised approaches.

Keywords: Image harmonization · dual-encoder single-decoder · multihead attention · adaptive kernel

1 Introduction

With the development of digital entertainment and computer vision [36, 53, 13], integrating material from different sources into a coherent and authentic image has become a worthwhile endeavor [42, 35, 37]. However, composite images created through simple matting and stitching often appear unrealistic due to varying shooting conditions, such as weather [25], illumination [37], camera filters [57], aperture [27] and *etc.*. Therefore, image harmonization, a technique aimed at enhancing the realism of composite images

* Both authors contributed equally to this research.
** Corresponding Author.



(a) Composite 1 (b) GT 1 (c) Composite 2 (d) GT 2

Fig. 1: Illustration of Foreground and Background Region Characteristics in Composite Images. Panel (a) and Panel (c) display composite images showing the foreground and specific regions marked for analysis. Panels (b) and (d) show the corresponding ground truth images. Color-coded boxes highlight regions with similar colors but differing textures between foreground and background, demonstrating the challenges in color harmonization.

[56, 60], is very important and challenging. In the realm of composite visuals, the most crucial problem is that there are discrepancies in appearance and semantics between source and target images [30].

As deep learning technologies have advanced, a variety of methods utilizing neural networks to harmonize composite images have emerged. Most of them implement end-to-end image harmonization networks through complex network designs, *i.e.*, RainNet [33] introduced Region-Aware Adaptive Instance Normalization (RAIN), which captures the style statistics information from the background features and applies it to the foreground. S^2AM[8] employs Spatial-Separated Attention Module to learn pixel-to-pixel translation.

Although these methods have achieved impressive results in image harmonization, they primarily focus on the color distribution of composite images and often neglect the intricate details critical to an image's perceptual quality. In particular, the nuanced variations in texture, which significantly influence an object's appearance [15, 44, 59], are not sufficiently addressed.

In a composite image, multiple background regions may share similar colors with the foreground. As shown in figure 1 (a) and (b), the foreground wooden planks in the composite image are brownish-yellow, aligning more closely in color with the background plants than the blue wooden planks in the background. Conversely, the real image reveals that the wooden planks should closely resemble those in the background. Another example is figure 1 (c) and (d), the foreground rock in the composite image is a bluish-gray color, more similar to the color of the sea surface, yet its texture is more similar to the rock wall in the background. Relying exclusively on color similarity for guiding color transformations between the foreground and background regions may lead to visual inconsistencies that undermine the overall realism of the composite image. This is because such an approach overlooks the crucial role of texture, which often conveys critical information about the material properties and contextual relevance of

the objects within the image. Incorporating texture matching between the foreground and background into the harmonization process can significantly enhance the effectiveness of the harmonization, ensuring that the visual output is not only cohesive but also more realistic.

To better utilize the texture and color features, we developed an innovative image harmonization network *TCDNet*. We utilize a designed dual-encoder single-decoder architecture to harmonize composite images through a unified texture-color perspective, with separate texture and color encoders to decouple and extract distinct features. With this dual-encoder single-decoder architecture, the network efficiently acquires decoupled texture and color features. Subsequently, we introduce the Texture Based Color Transfer (TBCT) module, which employs the extracted texture features to guide the color transfer process.

Considering that color and texture are naturally coupled, it is crucial to extract texture individually to utilize texture feature to guide the following color harmonization [59, 23]. Extracting texture independently from color without supervise, however, poses significant challenges, despite the use of a dedicated Texture Encoder. To address this issue, we proposed a texture loss that specifically disregards color information, concentrating solely on capturing accurate texture representations.

However, there might be some mismatched or unmatched regions between foreground and background [51]. So the harmonization of such areas will decrease with only the previous design.For these regions, we propose Color Dynamic (CoDy) module in the decoder, which utilize adaptive kernels for different inputs to transfer color between foreground and background.

With those designs mentioned above, the proposed framework is more effective compared to existing image harmonization models. The main contributions can be summarized as follows:

- We propose a novel color-texture dual-encoder single-decoder image harmonization network, which adaptively utilize texture to guide color transfer. To our knowledge, *TCDNet* is the first image harmonization models that consider the relationship between texture and color.
- We present a Texture Based Color Transfer (TBCT) module which divide the image into different regions, and find the region that match with the corresponding foreground region. This allows the model to effectively transfer color from the background to the foreground.
- We develop a Color Dynamic (CoDy) module, which generate adaptive kernels to learn the representations of foreground and background regions leading to better visual consistency.

2 Related Work

2.1 Image Harmonization

Image harmonization, as a subtask in image composition, aims to adjust the appearance of the foreground to make it consistent with background. Traditional image harmonization methods mainly depend on analyzing and manipulating low-level hand-crafted features, such as employing multi-scale various statistics [46] and gradient information[24,

40, 47]. There are also some traditional methods achieve visual consistency between the foreground and background through color transformation [41, 56].

With the advancement of deep learning, numerous deep learning-based methods have emerged, showing significant efficacy in image harmonization. Some of them utilize domain translation to enhance the consistency between foreground and background to harmonize images, such as DoveNet[7] and BargainNet[5]. Methods such as S2CRNet[32] employ color transformation into image harmonization. There are also many methods that treat image harmonization as a style transfer task. For instance, Ling et al.[33] proposed Region-aware Adaptive Instance Normalization (RAIN) module, an innovative approach that transfers the statistical properties of background features to normalized foreground features, demonstrating its adaptability and effectiveness in applications requiring localized adjustments to achieve desired visual effects.

However, some of the approaches mentioned above often overlook texture, which is a crucial aspect of an object's visual appearance. Additionally, some fail to account for the coupling between texture and color, as well as the influence of texture on color, potentially leading to unnatural visual effects. To address this oversight, our proposed TCDNet employs a unified texture-color perspective. By simultaneously adjusting both texture and color, our model enhances the perceptual quality of composite images and ensures visual consistency between the foreground and background elements in composite images.

2.2 Dynamic Neural Network

Dynamic neural networks [3, 16] have gained considerable attention in recent years for their ability to adapt their computational structure based on the input, which improves the efficiency and performance across various computer vision tasks. Unlike static architectures, dynamic neural networks can adjust their structures or parameters dynamically based on different inputs. Attention modules [54], as a common type of dynamic network, compute attention maps to highlight important channels or regions. However, adjusting the weights of each pixel individually may lead to a loss of the translational invariance inherent in convolutional neural networks (CNNs). Dynamic Region-aware Convolution [3] solves this problem by assigning multiple convolutional filters to different regions while sharing the same filters within each region.

As for computer vision, dynamic neural networks have been utilized for a variety of tasks, such as such as image classification [58, 20], semantic segmentation [31] and object detection [10]. In this work, we incorporate a dynamic approach within our proposed Color Dynamic module, which enables the adjustment of normalization parameters based on specific regional information. This adaptive capability enhances the harmonization results by ensuring more precise and context-aware color transformation.

2.3 Style Transfer

Style transfer is a technique in computer vision and image processing that applies the artistic style of one image to another, while retaining the content of the second image [26]. In recent years, this process often involves leveraging deep learning models to extract and transfer style features while preserving the content structure. Gatys et

al.[11] renders a content image in the style of another image, paving the way of subsequent research of this field. Ulyanov et al.[49] presents a feed-forward approach for texture synthesis and style transfer. Following these, Huang et al.[21] proposed Adaptive Instance Normalization (AdaIN) which adjusts the content image to mimic the style image's statistical features, allowing for effective style adaptation. Another approach, Interactive Image Style Transfer Network (IIST-Net)[52], utilizes an interactive brush-texture generation module and multilayer style attention to produce realistic artistic stylizations guided by user-defined graffiti curves. Whitening and Coloring Transform (WCT)[4], transforms content features to align with style features through a two-step process, enabling precise content-style amalgamation.

While style transfer modifies images to mimic artistic styles, image harmonization addresses a different challenge by seamlessly integrating the foreground and background of composite images to preserve their natural appearance. There are some style transfer methods that target realistic reference styles, but most style migration tasks focus primarily on artistic styles. However, image harmonization prioritizes consistency and realism among image components. The strong reliance on texture abstraction in style transfer renders it unsuitable for harmonizing real photographs. However, our TCDNet employs a texture-color perspective that effectively transfers background textures to the foreground while maintaining color correlations, enhancing the realism of synthetic images.

3 Method

3.1 Overview

Image harmonization refers to the process of adjusting an image's color, illumination, composition, or other visual elements during editing or creation, so that it aligns more closely with aesthetic standards and enhances overall visual comfort. In the context of composite image optimization, the aim of image harmonization is to adjust the appearance of the foreground I_f to make it compatible with the background I_b in a composite image I_c. Considering the image harmonization network as a generator G, the inputs of G are composite image I_c and the mask of its foreground M. The harmonized image will be generated as $\hat{I} = G(I_c, M)$. Our goal is to improve the harmonization effect for making close to the ground truth image by minimizing $\|G(I_c, M) - I\|$.

In order to enhance the realism of harmonized composite image, we proposed TCDNet. The overall network architecture of TCDNet, as shown in Figure 2, comprises two encoders (a Texture Encoder (TE) and a Color Encoder (CE)), a Texture Based Color Transfer (TBCT) module, and a Decoder with Color Dynamic (CoDy) module. These components are detailed in the following subsection.

3.2 Dual Encoder

Most image harmonization tasks only focus on enhancing color harmony between the foreground and background by analyzing color characteristics. However, both color and texture features play an essential role in image harmonization. As shown in Figure 5,

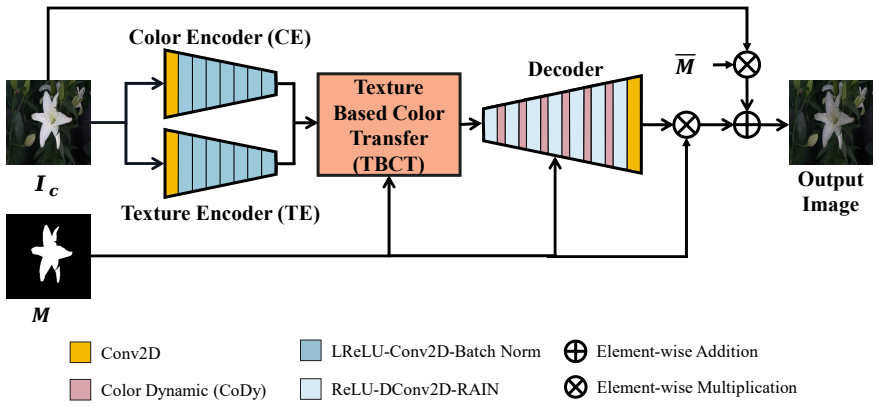


Fig. 2: The overview of our proposed Texture and Color Dynamic Network. TCDNet consist of dual encoder, a Texture Based Color Transfer module, and a decoder with several Color Dynamic modules. The dual encoder consists of a color encoder and a texture encoder, each designed to independently extract color and texture features without influencing each other.

even though the swans in the first row of images are grayish-white, the texture of their feathers causes distinct visual patterns across different areas of their wings. Neglecting the influence of texture in visual impact of images can lead to the loss of texture details in the harmonized images [38]. Additionally, analyzing texture enhances the model's ability to identify and recognize the same visual pattern among the objects from foreground and background, enabling it to transfer relevant appearance from background to foreground if they have the same texture. To effectively leverage both texture and color information, our framework employs the dual-encoder consisting of separate texture and color encoders, each specialized in extracting its respective feature set.

As depicted in Figure 2, each encoder within our network is structured with a two-dimensional convolutional layer followed by seven layers that each combine Leaky ReLU activation layer[34], convolution layer, and batch normalization layer [22]. Each combined layer l, represented as Leaky ReLU-Conv2D-Batch Norm, consists of a Leaky ReLU activation function followed by a Conv2D layer, and is then followed by batch normalization layer. The color encoder and texture encoder shares the same hyper parameters for network architecture. For detail, we first encode the input image with shape $(3, 256, 256)$ into a primary feature map of shape $(32, 256, 256)$. After the whole encoder, the color feature and texture feature are of shape $(256, 32, 32)$.

The primary function of the texture encoder within our network is to isolate and extract texture features independently of other image features such as color. This is crucial because texture plays a significant role in achieving realistic image harmonization. However, texture and color features are inherently coupled [18], making it difficult to extract separate texture features that are not affected by color. As a result, traditional extraction techniques may inadvertently capture color characteristics when isolating texture, thereby confounding the texture analysis. The inherent coupling between chro-

matic information and textural patterns poses a fundamental challenge for their independent extraction. However, through comparative analysis of color images and their grayscale versions, it is demonstrable that textural features remain consistent regardless of color variations. This perceptual invariance suggests that robust texture feature extraction can be achieved by identifying invariant patterns within multi-chromatic representations of identical visual content, thereby enabling the derivation of extraction of texture feature through color-independent feature analysis feature analysis. Therefore, we design a texture loss as shown in Equation. 1 to pull in the texture feature distribution of real image and composite image.

$$L_{texture} = MSE(TE(I_{real}), TE(I_{comp})) \tag{1}$$

This loss function is designed to refine the texture feature extraction process by comparing the texture distribution between the real and composite images, where TE donates extractor texture features from given image, that is, Texture Encoder. By focusing on minimizing this texture-specific loss, we ensure that the texture features we extract are not influenced by the color variations inherent to the images. This method not only reinforces the independence of texture features but also significantly enhances the quality of the harmonization by ensuring that texture adjustments are made purely based on texture discrepancies, rather than being affected by color.

3.3 Texture Based Color Transfer Module

The dual encoder architecture significantly enhances our model's ability to independently extract texture and color features, each decoupled from the other, which is crucial for the following sophisticated image harmonization tasks. However, simply extracting these features independently does not inherently resolve the challenges associated with integrating the foreground more naturally into the background. To address this limitation, we developed the Texture Based Color Transfer (TBCT) module as Figure 3, which utilizes texture-driven guidance to facilitate color transfer and improve image harmonization. We employ attention mechanism [1] within TBCT to identify regions in the background that share textural similarities with the foreground pixels. By identifying regions within the background that exhibit textural similarities to those in the foreground, TBCT can more effectively guide the transfer of color based on texture matching, thereby enhancing the overall harmony and realism of the composite image.

Specifically, after processing each image through the dual encoder, we can obtain its texture feature map $F_t \in \mathbb{R}^{C\times H\times W}$ and color feature map $F_c \in \mathbb{R}^{C\times H\times W}$. Additionally, we can also acquire the corresponding mask, $M \in \mathbb{R}^{C\times H\times W}$ from the input, where C, H, W indicate the number of channels, height, and width of F, respectively. The encoder feature F_t and F_c can be viewed as a set of $C \times H \times W$-dimensional local representations. By utilizing the mapping relationship of the mask, these representations can also be divided into foreground and background,

$$\begin{aligned} F_{tf} = F_t \times M, F_{tb} = F_t \times (1 - M) \\ F_{cf} = F_c \times M, F_{cb} = F_c \times (1 - M) \end{aligned} \tag{2}$$

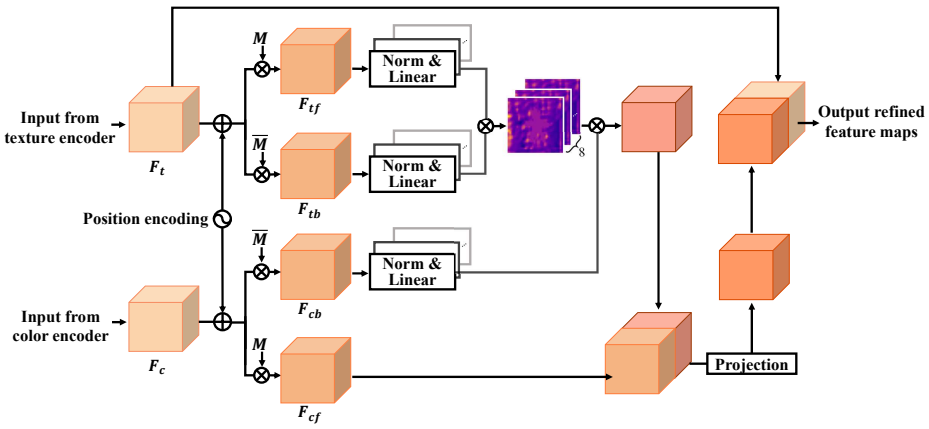


Fig. 3: The architecture of Texture Based Color Transfer Module. Our TBCT module identifies regions in the background that share textural similarities with the foreground pixels, and leverage this texture as guidance for color transfer.

where $F_{tf}, F_{cf} \in \mathbb{R}^{C\times N_f}$, $F_{tb}, F_{cb} \in \mathbb{R}^{C\times N_b}$ are the foreground and background feature maps of the texture and color feature, respectively. N_f is the number of foreground representations, and N_b is the number of background representations, $N_f + N_b = HW$. The feature maps can be denoted as:

$$\begin{aligned} Q &= Norm(F_{tf}) \\ K &= Norm(F_{tb}) \\ V &= Norm(F_{cb}) \end{aligned} \tag{3}$$

Since a composite image divided into foreground and background by a mask, our goal is to identify the background representation that most closely matches the texture of each foreground segment. To achieve this, we employ a multi-head attention mechanism [50] within a texture attention block, enabling the network to selectively transfer relevant appearance attributes when textural consistency exists between the foreground and background. Thus, the multi-head attention can be calculated as:

$$Multihead(Q, K, V) = Concat(head_1, head_2, ..., head_h) \tag{4}$$

$$head_i = Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

we employ $h = 8$ parallel heads, and the dimension $D_k = 512$.

Considering the importance of maintaining spatial information, especially as representations may lose their positional context during division processes. To address this, we integrate position encoding [9] within TBCT. Position encoding embeds implicit spatial features into the model, enhancing its ability to maintain the geometric and spatial continuity essential for visual coherence. This not only enhances the model's ability to discern subtle spatial details but also promotes more effective feature interactions

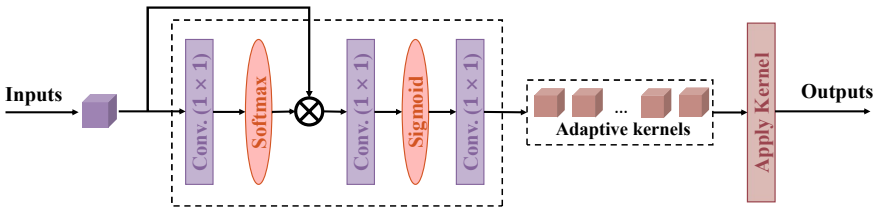


Fig. 4: The network architecture of CoDy. Our CoDy modules are inserted between decoder layers, enabling the network to process color information from the global feature space and predict adaptive convolutional kernels for seamless foreground-background blending.

across different regions of the image, ultimately leading to improved harmonization outcomes.

3.4 Decoder with Color Dynamic Module

Similar to the encoder, the decoder consists of seven combined layers. As depicted in Figure 2, the decoder's architecture employs a ReLU-DeConv.-RAIN configuration, in contrast to the encoder's use of Leaky ReLU. This shift to ReLU enhances computational efficiency and stability by focusing on non-negative activations, which is critical for effective image reconstruction. Leaky ReLU, used in the encoder, helps preserve information during deep neural network processing by allowing a small, non-zero gradient when the unit is otherwise inactive. Furthermore, we adopt RAIN [33] for normalization within the decoder. RAIN, or Region-Aware Adaptive Instance Normalization, dynamically adjusting instance normalization parameters based on region-specific information, enhancing the harmonization of composite images by accounting for varying styles across different regions.

Although TBCT utilizes the texture matching relationship between the foreground and background to guide color transfer and harmonize the composite image, there may be some weak texture regions in a composite image. These regions may affect the effectiveness of texture feature extraction, limiting TBCT's ability to achieve seamless harmonization. Additionally, some foreground regions may not find matching areas with similar textures in the background. Exclusively relying on TBCT may not effectively harmonize these regions. Therefore, we propose the Color Dynamic Module (CoDy), which transfers color between mismatched or unmatched regions in the foreground and background using adaptive kernels [45]. By employing adaptive kernels, the CoDy module dynamically adjusts its parameters based on the specific features of each input region, facilitating precise and context-aware color correction that ensures the foreground seamlessly integrates into the background. We append a CoDy module to each combined layer as shown in Figure 2.

Figure 4 shows the network architecture of CoDy. This module enables the network to utilize color information from the global feature space and to predict adaptive convolutional kernels. The adaptive kernels are generated and applied in a context-aware

Table 1: Quantitative comparison of several state-of-the-art image harmonization models. Top two performance are shown in **red** and **blue**. ↓ means the lower the better, and ↑ means the higher the better.

Model	HAdobe5k		HFlickr		HCOCO		Hday2night		Average	
	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑
Composite	345.54	28.16	264.35	28.32	69.37	33.94	109.65	34.01	172.47	31.63
DIH [48]	92.65	32.28	163.38	29.55	51.85	34.69	82.34	34.62	76.77	33.41
S^2AM [8]	63.40	33.77	143.45	30.03	41.07	35.47	76.61	34.50	59.67	34.35
iS^2AM [43]	21.60	38.28	69.43	33.65	16.15	39.40	40.39	37.87	24.13	38.41
DoveNet [7]	52.32	34.34	133.14	30.21	36.72	35.83	51.95	35.27	52.33	34.76
RainNet [33]	43.35	36.22	110.59	31.64	29.52	37.08	57.40	34.83	40.29	36.12
BargainNet [5]	39.94	37.92	97.32	31.34	24.84	37.03	50.98	35.67	37.82	35.88
Intrinsic [14]	43.02	35.20	105.13	31.34	24.97	37.16	55.53	35.96	38.71	35.90
Harmonizer [28]	21.89	37.64	64.81	33.63	17.34	38.77	33.14	37.56	24.26	37.84
S^2CRNet [32]	34.91	36.42	98.73	32.48	23.22	38.48	51.67	36.81	35.58	37.18
DCCF [55]	23.34	37.75	64.77	33.60	17.07	38.66	55.76	37.40	24.65	37.87
CDTNet [6]	**20.62**	38.24	68.61	33.55	16.25	39.15	37.92	37.95	23.75	37.85
SCS-Co [17]	21.01	**39.21**	**55.83**	**34.22**	**13.58**	**41.75**	37.83	**38.41**	**21.33**	38.75
SP-IC cycle [2]	**18.17**	38.91	68.85	33.88	14.82	39.73	**31.47**	37.90	22.47	**38.81**
TCDNet(ours)	20.80	**39.05**	**53.75**	**34.72**	**13.30**	**40.16**	**32.10**	**38.36**	**20.35**	**39.20**

fusion, ensuring accurate capture of the color correlation between the foreground and background. The number of adaptive kernels is $2 * c^2$, which c is the number of input channels. Moreover, we employ the mask to provide spatial guidance to the network, indicating the regions that require harmonization versus those that serve as reference.

4 Experiments

In this section, we first introduce the datasets, metrics, and implementation details utilized in our experiments. Subsequently, we compare the performance of TCDNet against other image harmonization methods. We also conduct ablation studies to assess the impact of each module within TCDNet, providing insights into the observed performance improvements. Lastly, we explore the effects of various hyperparameter selections on our results.

4.1 Experiment Settings

Datasets. We choose the iHarmony4 benchmark[7] for the training and evaluation for TCDNet. iHarmony4 is a widely used benchmark dataset in image harmonization, which consists of 73,146 image pairs and includes four subsets: HAdobe5k, HFlickr, HCOCO, and Hday2night. For each sample in iHarmony4, there is a natural image as ground truth, a foreground mask, and a composite image (with the foreground generated by GAN[12]). In this paper we use filtered synthetic images from the iHarmony4

Table 2: The mean square error of the foreground region (fMSE). Top two performance are shown in **red** and **blue**. The lower value of fMSE means the better.

Model	0% ∼ 5%	5% ∼ 15%	15% ∼ 100%	Average
Composite	1208.86	1323.23	1887.05	1387.30
S^2AM [48]	509.41	454.21	449.81	481.79
DoveNet [7]	591.88	504.42	505.82	549.96
RainNet [33]	550.38	378.69	389.80	469.60
BargainNet [5]	450.33	359.49	353.84	405.23
Intrinsic [14]	441.02	363.61	354.84	400.29
S^2CRNet [32]	**239.94**	271.70	333.96	274.99
SP-IC cycle [2]	276.59	**209.56**	**216.37**	**245.75**
TCDNet(ours)	**259.50**	**199.03**	**193.46**	**229.36**

dataset as 'ground truth' for evaluation. While iHarmony4 attempts to remove poorly synthesized samples, there is no guarantee that these images are objectively perfect or optimal; rather, they are the generally accepted industry reference standard. We follow the same partition settings of iHarmony4 as DoveNet[7].

Implementation Details. The proposed TCDNet is implemented with PyTorch[39] with Nvidia RTX A4500 GPU. We utilize Adam optimizer [29] and set $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batchsize is 16, and we train the TCDNet for 120 epochs. The initial learning rate is 0.001. It starts linearly decrease at 30-th epoch and decreases to 0 at 120-th epoch. All images are resized to 256×256 with no data augmentation adopted.

Evaluation Metrics. We assess the performance using a set of commonly used metrics: Mean Square Error (MSE), foreground MSE (fMSE), and Peak Signal-to-Noise Ratio (PSNR) [19]. Additionally, we provide qualitative comparisons with various state-of-the-art methods to illustrate the performance of our approach.

4.2 Comparison with Other Methods

Performances of different sub-datasets. Table 1 presents the quantitative results of previous state-of-the-art methods and our *TCDNet* on different subset. From Table 1, we can observe that, *(i)* Our method demonstrates exceptional performance across all metrics and subsets with average MSE of our model is 20.35 and PSNR is 39.20, which demonstrates that *TCDNet* can achieves superior and stable performance on image harmonization. *(ii) TCDNet* particularly excels in PSNR. This is attributed to the model's ability to maintain the overall image structure and quality, effectively preserving both high-frequency details and low-frequency components, which contributes to its superior performance in various quantitative evaluations. *(iii)* The most notable performance improvement is observed in the HFlickr subset. This improvement is likely due to the higher quality of images in HFlickr, which are typically shot by photographers and exhibit better color and texture conditions, thus responding exceptionally well to our model.

Influence of foreground ratios. Following [33], we tested the effect of different foreground ratios on the effect of network harmonization. The foreground ratio refers to the

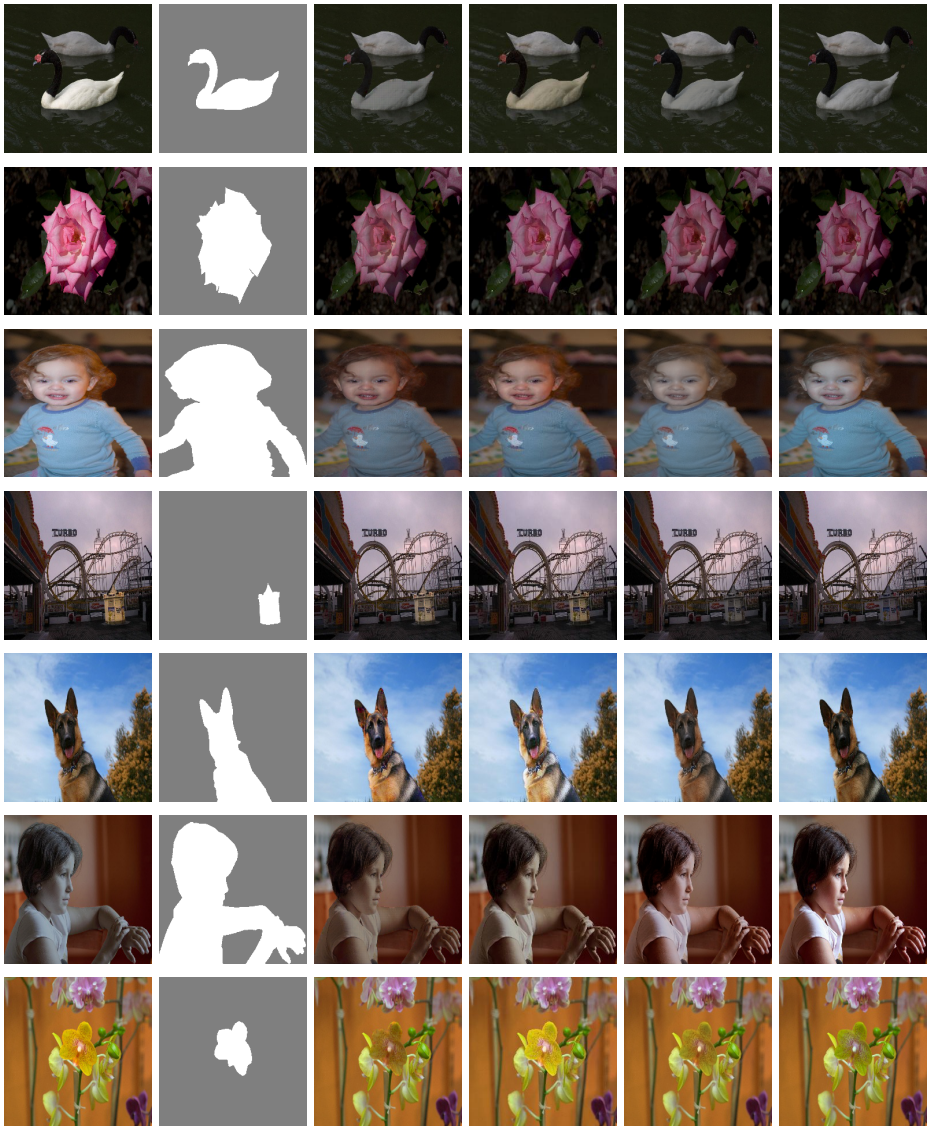


Fig. 5: Comparison of different methods. Our method achieves color harmony while preserving texture details, resulting in more photo-realistic outcomes.

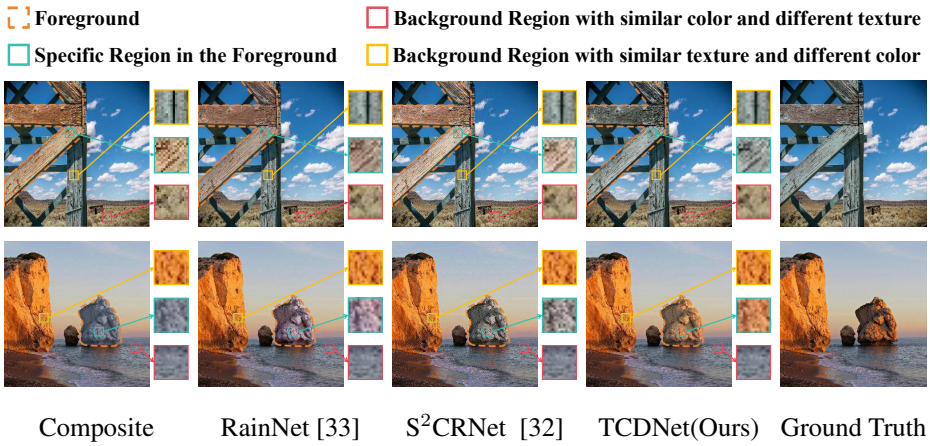


Fig. 6: Comparison of regional harmonization effects. Our TCDNet accounts for both texture and color matching between the foreground and background, resulting in more realistic harmonization outcomes.

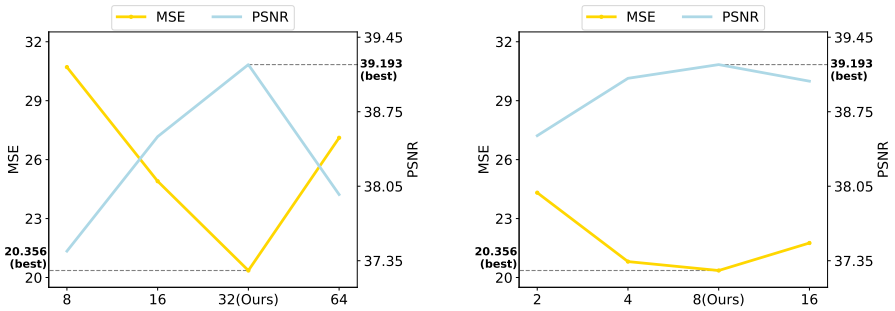
percentage of the foreground region in the overall area of the composite image, and we divide the ratios into three different buckets, i.e., 0% to 5%, 5% to 15%, 15% to 100%. The results of are shown in Table 2. We can find that, *(i) TCDNet* shows significant improvement in the 15% to 100% range, with a growth of 22.91 in fMSE. That's mainly because the CoDy module gives out more global and effective optimization of large foreground areas. *(ii)* The 0% to 5% ratio part of get a improvement of 17.09 in fMSE. This is because these foreground areas usually have background areas with similar textures, so the TBCT module can help them migrate and optimize color expression.

Qualitative comparisons. Figure 5 shows the qualitative comparison results between TCDNet and other methods,demonstrating that our TCDNet achieves color harmony while preserving texture details, resulting in better visual consistency and more photorealism. We can also observe that, *(i)* Of the first row of figure 5, RainNet [33] struggles to accurately reconstruct the fine textures of the swan's wings, whereas *TCDNet* successfully preserves the detailed feather texture on the wings and resulting in better visual outcomes. This is mainly due to the dual encoder structure that introduces texture information that has been neglected in previous work and adds texture information to the reconstruction process of the harmonized image, reducing information loss. *(ii)* The fourth row of figure 5 provides an example of a serious performance difference, where RainNet [33] and S^2CRNet [32] could hardly complete the task with the output image more similar to composite image. However, *TCDNet* achieve superior performance in this case. We believe this is mainly due to the fact that our model relies on texture rather than color to guide the learning of color from the background, enabling it to still achieve such performance in such a complex environment.

Furthermore, we compared the results of the detailed region between *TCDNet* and other methods as shown in Figure 6. *TCDNet* consistently excels in harmonizing foreground regions that exhibit similar colors but differing textures from the background We could also notice that, *(i)* RainNet [33] and S^2CRNet [32] give out a region with

Table 3: Ablation study comparing the fMSE of our TCDNet. The lower the better.

Model	$0\% \sim 5\%$	$5\% \sim 15\%$	$15\% \sim 100\%$	Average
Ours w/o TE	386.27	263.35	260.77	326.90
Ours w/o TBCT	317.50	231.03	241.04	278.31
Ours w/o CoDy	299.28	244.05	305.16	286.16
Ours	**259.50**	**199.03**	**193.46**	**229.36**



(a) Number of output channels in first layer encoder.

(b) Number of heads in multihead attention.

Fig. 7: Influence of hyper-parameters.

color similar to background region with similar color and different texture, which confirms that these methods, which mainly focus on the transfer of similar color regions, are not effective in achieving harmonization in the face of such situations. *(ii) TCDNet* is the opposite, in that it harmonizes foreground regions with similar texture regions in their background, which is also close to the ground truth. This shows that our migration method focusing on similar texture regions can effectively achieve color harmony in such regions and demonstrates the effectiveness of *TCDNet*.

4.3 Ablation Study

The effectiveness of Dual Encoder. The dual-encoder design enables our network to decouple color and texture of the image, facilitating the independent extraction of these features. As shown in Table 3, removing either encoder significantly degrades the network's performance, with the average fMSE increasing by 97.54. This is mainly due to the dual-encoder architecture introduces a whole new level of information about the image, allowing the model to focus on more information in the image.

The effectiveness of Texture Based Color Transfer Module. Our Texture Based Color Transfer Module matches regions with similar textures between the foreground and background using an attention mechanism and leverages texture-driven color transfer. According to Table 3, the removal of TBCT from TCDNet results in a significant incre ase in the fMSE for the 0% to 5% foreground range and slight increase on other range; a lower fMSE indicates better performance. Considering that images in the 0% to 5% foreground range usually have areas with similar textures to the foreground, such as

multiple people, multiple flowers, *etc.*. This bias of performance degradation within different range also confirms that the design of Texture Based Color Transfer Module allows the model to learn color information from areas with similar textures, thereby bringing better performance.

The effectiveness of Color Dynamic Module. Our Color Dynamic (CoDy) Module utilizes adaptive kernels to transfer color between mismatched or unmatched regions in the foreground and background. Results in Table 3 indicate that models with CoDy degenerate greatly in foreground range of 15% to 100%, with a 111.70 difference than ours model. In the meantime, the difference of other range is less than 45. This is because the foregrounds with foreground range greater than 15% occupy a considerable portion of the image area, so there are usually no areas in the image with similar textures, therefore TBCT module cannot find similar texture area and could not effectively harmonize the image. This confirms that adaptive kernels in CoDy module are essential for effectively addressing color discrepancies between the foreground and background, thus improving the visual quality of the harmonized images.

The selection of parameters. In the training process of TCDNet, there are some hyper-parameters are involved. We conducted experiments to determine the optimal values for these parameters. The experimental results are shown in Figure 7, from which we determined that the optimal values for multihead and the number of output channels in first layer encoder are 32 and 8, respectively.

5 Conclusion

This paper proposes a novel network that realize image harmonization in a texture-color perspective. Dual-encoder extract separate texture and color features for following harmonization procedure. Texture-Based Color Transfer module leverage attention mechanism to find the regions with similar texture between foreground and background, and to guide the color transfer by texture. The decoder reconstructs harmonized images using concatenated texture-color features. The Color Dynamic module within decoder employs adaptive kernels to optimize features in regions with weak textures. Our method achieves outstanding performances on the iHarmony4 benchmark dataset. The primary limitation stems from the TBCT module's performance degradation in areas with weak textures. Future work will continue to explore this direction.

Acknowledgments. This work was supported by the National Key R&D Program of China under Grant 2022YFF0904300.

References

1. Bahdanau, D.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
2. Cai, X., Shi, Q., Gao, Y., Li, S., Hua, W., Xie, T.: A structure-preserving and illumination-consistent cycle framework for image harmonization. IEEE Transactions on Multimedia **26**, 51–64 (2023)

3. Chen, J., Wang, X., Guo, Z., Zhang, X., Sun, J.: Dynamic region-aware convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8064–8073 (2021)
4. Chiu, T.Y.: Understanding generalized whitening and coloring transform for universal style transfer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4452–4460 (2019)
5. Cong, W., Niu, L., Zhang, J., Liang, J., Zhang, L.: Bargainnet: Background-guided domain translation for image harmonization. In: 2021 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2021)
6. Cong, W., Tao, X., Niu, L., Liang, J., Gao, X., Sun, Q., Zhang, L.: High-resolution image harmonization via collaborative dual transformations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18470–18479 (2022)
7. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8394–8403 (2020)
8. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. IEEE Transactions on Image Processing **29**, 4759–4771 (2020)
9. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Figurnov, M., Collins, M.D., Zhu, Y., Zhang, L., Huang, J., Vetrov, D., Salakhutdinov, R.: Spatially adaptive computation time for residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1039–1048 (2017)
11. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2414–2423 (2016)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
13. Guo, J., Zhang, D., Liu, X., Zhong, Z., Zhang, Y., Wan, P., Zhang, D.: Liveportrait: Efficient portrait animation with stitching and retargeting control. CoRR (2024)
14. Guo, Z., Zheng, H., Jiang, Y., Gu, Z., Zheng, B.: Intrinsic image harmonization. In: Proceedings of the ieee/cvf conference on computer vision and pattern recognition. pp. 16367–16376 (2021)
15. Haindl, M., Filip, J.: Visual texture: Accurate material appearance measurement, representation and modeling. Springer Science & Business Media (2013)
16. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., Wang, Y.: Dynamic neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 7436–7456 (2021)
17. Hang, Y., Xia, B., Yang, W., Liao, Q.: Scs-co: Self-consistent style contrastive learning for image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19710–19719 (2022)
18. Hoang, M.A., Geusebroek, J.M., et al.: Measurement of color texture. In: Workshop on Texture Analysis in Machine Vision. pp. 73–76. Citeseer (2002)
19. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
20. Huang, G., Chen, D., Li, T., Wu, F., Van Der Maaten, L., Weinberger, K.Q.: Multi-scale dense networks for resource efficient image classification. ICLR (2018)
21. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. pp. 1501–1510 (2017)

22. Ioffe, S.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
23. Ji, G.P., Fan, D.P., Chou, Y.C., Dai, D., Liniger, A., Van Gool, L.: Deep gradient learning for efficient camouflaged object detection. Machine Intelligence Research **20**(1), 92–108 (2023)
24. Jia, J., Sun, J., Tang, C.K., Shum, H.Y.: Drag-and-drop pasting. ACM Transactions on graphics (TOG) **25**(3), 631–637 (2006)
25. Jiang, Z., Zhang, Z., Fan, X., Liu, R.: Towards all weather and unobstructed multi-spectral image stitching: Algorithm and benchmark. In: Proceedings of the 30th ACM international conference on multimedia. pp. 3783–3791 (2022)
26. Jing, Y., Yang, Y., Feng, Z., Ye, J., Yu, Y., Song, M.: Neural style transfer: A review. IEEE transactions on visualization and computer graphics **26**(11), 3365–3385 (2019)
27. Kaur, H., Koundal, D., Kadyan, V.: Image fusion techniques: a survey. Archives of computational methods in Engineering **28**(7), 4425–4447 (2021)
28. Ke, Z., Sun, C., Zhu, L., Xu, K., Lau, R.W.: Harmonizer: Learning to perform white-box image and video harmonization. In: European Conference on Computer Vision. pp. 690–706. Springer (2022)
29. Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
30. Li, A., Guo, J., Guo, Y.: Image stitching based on semantic planar region consensus. IEEE Transactions on Image Processing **30**, 5545–5558 (2021)
31. Li, X., Liu, Z., Luo, P., Change Loy, C., Tang, X.: Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3193–3202 (2017)
32. Liang, J., Cun, X., Pun, C.M., Wang, J.: Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In: European Conference on Computer Vision. pp. 334–349. Springer (2022)
33. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9361–9370 (2021)
34. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: Proc. icml. vol. 30, p. 3. Atlanta, GA (2013)
35. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
36. Nalbant, K.G., Uyanık, Ş.: Computer vision in the metaverse. Journal of Metaverse **1**(1), 9–12 (2021)
37. Nie, L., Lin, C., Liao, K., Liu, S., Zhao, Y.: Unsupervised deep image stitching: Reconstructing stitched features to images. IEEE Transactions on Image Processing **30**, 6184–6197 (2021)
38. Olkkonen, M., Hansen, T., Gegenfurtner, K.R.: Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. Journal of vision **8**(5), 13–13 (2008)
39. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
40. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 577–582 (2023)
41. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer graphics and applications **21**(5), 34–41 (2001)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

43. Sofiiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1620–1629 (2021)
44. Song, Z.C., Liu, S.G.: Sufficient image appearance transfer combining color and texture. IEEE Transactions on Multimedia **19**(4), 702–711 (2016)
45. Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., Kautz, J.: Pixel-adaptive convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11166–11175 (2019)
46. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Transactions on Graphics (TOG) **29**(4), 1–10 (2010)
47. Tao, M.W., Johnson, M.K., Paris, S.: Error-tolerant image compositing. International journal of computer vision **103**, 178–189 (2013)
48. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3789–3797 (2017)
49. Ulyanov, D., Lebedev, V., Lempitsky, V., et al.: Texture networks: Feed-forward synthesis of textures and stylized images. In: International Conference on Machine Learning. pp. 1349–1357. PMLR (2016)
50. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
51. Wang, K., Gharbi, M., Zhang, H., Xia, Z., Shechtman, E.: Semi-supervised parametric real-world image harmonization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5927–5936 (2023)
52. Wang, Q., Ren, Y., Zhang, X., Feng, G.: Interactive image style transfer guided by graffiti. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 6685–6694 (2023)
53. Ward, T.M., Mascagni, P., Ban, Y., Rosman, G., Padoy, N., Meireles, O., Hashimoto, D.A.: Computer vision in surgery. Surgery **169**(5), 1253–1256 (2021)
54. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
55. Xue, B., Ran, S., Chen, Q., Jia, R., Zhao, B., Tang, X.: Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In: European Conference on Computer Vision. pp. 300–316. Springer (2022)
56. Xue, S., Agarwala, A., Dorsey, J., Rushmeier, H.: Understanding and improving the realism of image composites. ACM Transactions on graphics (TOG) **31**(4), 1–10 (2012)
57. Yan, N., Mei, Y., Xu, L., Yu, H., Sun, B., Wang, Z., Chen, Y.: Deep learning on image stitching with multi-viewpoint images: A survey. Neural Processing Letters **55**(4), 3863–3898 (2023)
58. Yang, L., Han, Y., Chen, X., Song, S., Dai, J., Huang, G.: Resolution adaptive networks for efficient inference. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2369–2378 (2020)
59. Zhang, H., Xue, J., Dana, K.: Deep ten: Texture encoding network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 708–717 (2017)
60. Zhu, J.Y., Krahenbuhl, P., Shechtman, E., Efros, A.A.: Learning a discriminative model for the perception of realism in composite images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3943–3951 (2015)