Hierarchical Dynamic Image Harmonization

Haoxing Chen
Nanjing University
Tiansuan Lab, Ant Group Inc.
Hangzhou, China
hx.chen@hotmail.com

Zhangxuan Gu
Tiansuan Lab, Ant Group Inc.
Shanghai, China
guzhangxuan.gzx@antgroup.com

Yaohui Li
NLP Research Group, Nanjing
University
Nanjing, China
yaohuili@smail.nju.edu.cn

Jun Lan
Tiansuan Lab, Ant Group Inc.
Hangzhou, China
yelan.lj@antgroup.com

Changhua Meng
Tiansuan Lab, Ant Group Inc.
Hangzhou, China
changhua.mch@antgroup.com

Weiqiang Wang
Tiansuan Lab, Ant Group Inc.
Hangzhou, China
weiqiang.wwq@antgroup.com

Huaxiong Li
Department of Control Science and
Intelligence Engineering, Nanjing
University
Nanjing, China
huaxiongli@nju.edu.cn

arXiv:submit/4876054 [cs.CV] 4 May 2023

ABSTRACT

Image harmonization is a critical task in computer vision, which aims to adjust the foreground to make it compatible with the background. Recent works mainly focus on using global transformations (i.e., normalization and color curve rendering) to achieve visual consistency. However, these models ignore local visual consistency and their huge model sizes limit their harmonization ability on edge devices. In this paper, we propose a hierarchical dynamic network (HDNet) to adapt features from local to global view for better feature transformation in efficient image harmonization. Inspired by the success of various dynamic models, local dynamic (LD) module and mask-aware global dynamic (MGD) module are proposed in this paper. Specifically, LD matches local representations between the foreground and background regions based on semantic similarities, then adaptively adjust every foreground local representation according to the appearance of its K-nearest neighbor background regions. In this way, LD can produce more realistic images at a more fine-grained level, and simultaneously enjoy the characteristic of semantic alignment. The MGD effectively applies distinct convolution to the foreground and background, learning the representations of foreground and background regions as well as their correlations to the global harmonization, facilitating local visual consistency for the images much more efficiently. Experimental results demonstrate that the proposed HDNet significantly reduces the total model parameters by more than 80% compared to previous methods, while still attaining state-of-the-art performance on the popular iHarmony4 dataset. Notably, the HDNet achieves a 4% improvement in PSNR and a 19% reduction in MSE compared to the prior state-of-the-art methods. Additionally, we introduced a lightweight version of HDNet, i.e., HDNet-lite, which has only 0.65MB parameters, yet it still achieved state-of-the-art performance. Our code is avaliable in https://github.com/chenhaoxing/HDNet.

KEYWORDS

Image harmonization, Hierarchical dynamics, K-nearest neighbor

1 INTRODUCTION

In the computer vision community, integrating image patches from different sources into a cohesive, realistic image is a foundational technique, as demonstrated in image editing [22, 29] and scene completion [2, 32]. However, composite images inevitably encounter inharmonious elements, as foreground and background appearances may have gaps due to varying imaging conditions (e.g., rainy and sunny, morning and dusk). Consequently, image harmonization, which strives for visual consistency within the composite image, constitutes an important and challenging task.

Traditional approaches emphasize the importance of harmonizing hand-crafted low-level appearance statistics, including color statistics [28, 36], and gradient information [24]. Nevertheless, these methods are unsatisfactory in the complex scenarios where the source image exhibits a significant appearance or semantic gap with the target.

With the advances in deep learning, more deep neural network-based methods were proposed [6–8, 18]. Most of them use complex network structures or training strategies [6, 8] to accomplish image harmonization tasks. In contrast, color transformation and normalization based models have received extensive attention due to their simplicity and flexibility [7, 21, 23].

Color transformation based techniques tend to learn an RGB-to-RGB transformation for image harmonization. For instance, Collaborative Dual Transformation (CDTNet)[7] incorporates a low-resolution generator for pixel-to-pixel transformation, lookup tables (LUTs) for RGB-to-RGB transformation, and a refinement module that takes advantage of both pixel-to-pixel transformation and RGB-to-RGB transformation.Recently, Spatial-Separated Curve Rendering Network (S^{2}CRNet)[21] introduces a curve rendering module (CRM) that learns and integrates spatial-specific knowledge, to generate piecewise curve mapping parameters in the foreground region.

Normalization based methods regard image harmonization as a background-to-foreground style transfer task. Inspired by AdaIN [16],
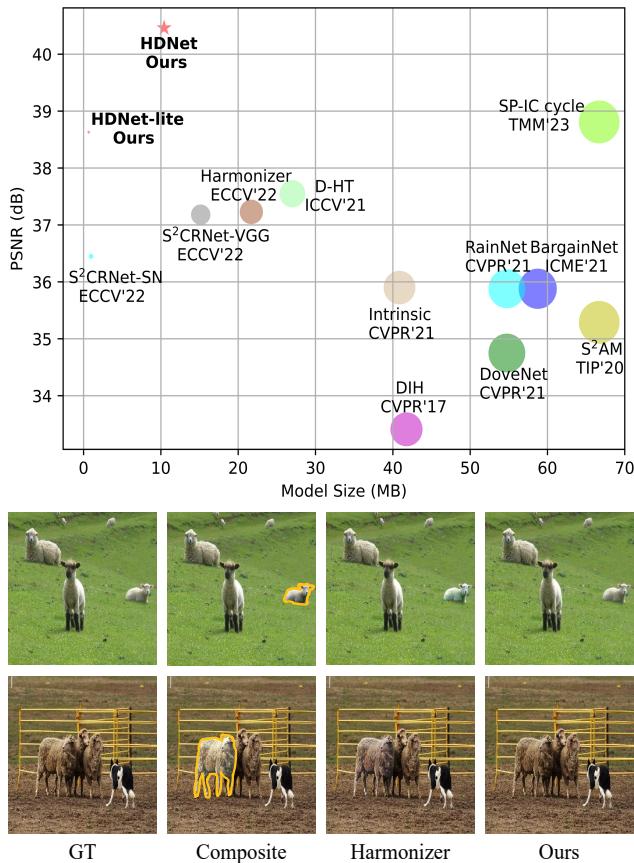
Figure 1: In the top figure, we compare parameter size and performance between our method and other state-of-the-art methods. It can be seen that our method has fewer parameters but achieve state-of-the-art results. In the bottom figure, our method produces a more photorealistic harmonized result.

Ling *et al.* [23] regard image harmonization as a background to foreground style transfer problem and proposed region-aware adaptive instance normalization (RAIN) which captures the style statistics information from the background features and applies it to the foreground. However, as shown in Figure 1, unexpected patterns still exist and are very severe in some cases.

However, global feature representation and transformation may not be effective enough for image harmonization tasks. This is because different regions of an image have distinct colors, textures, and structural characteristics that cannot be fully captured by global-level features. Another weakness of the above methods is that they use fixed statistics for normalization, which significantly limits their representation ability. Moreover, their model sizes are too large for edge devices, e.g., mobile phones.

In this paper, we solve the above problems by proposing an efficient dynamic image harmonization network, which hierarchically adapts the features by two dynamics, *i.e.*, local dynamic module and mask-aware global dynamic module from local to global view.

To align each foreground local representation with semantically and appearance matched to the background ones, local dynamic module first finds K-nearest neighbor background local representations. Then, local dynamic module adaptively reconstruct each foreground local representation by linearly combining it with related background local representations. Besides, for global feature learning, mask-aware global dynamic module utilizes distinct convs for both foreground and background regions. This module enables the model to acquire adaptive representations for these regions and effectively capture their correlations.

As shown in Figure 1, the proposed framework is efficient and effective compared to existing image harmonization models. Our method achieves higher performance with fewer parameters.

The main contributions can be summarized as follows:

- We propose a novel hierarchical dynamic image harmonization network, which adaptively adapt the features from local to the global view for background and foreground visual alignments.
- We present a local dynamic module, which finds K-nearest neighbor background local representations for each foreground local representation and adjusts the appearance of each foreground local representation.
- We develop a mask-aware global dynamic module to learn the representations of foreground and background regions as well as their correlations for the global harmonization, leading to better and more efficiently visual consistency.
- Evaluations on image harmonization datasets demonstrate that our method can achieve state-of-the-art performance using fewer parameters and lower computational costs.

2 RELATED WORKS

2.1 Image Harmonization

Traditional image harmonization methods primarily focus on enhancing composite images by manipulating low-level appearance features. These methods include color transformations to achieve a consistent visual appearance between foreground and background objects [28, 36], as well as employing multi-scale transformations and statistical analysis to adjust the overall composition [31]. Although these methods have shown satisfactory results in certain scenarios, they are limited by their reliance on low-level features and may struggle to adapt to more complex situations.

In recent years, deep learning-based approaches have emerged as a powerful alternative for image harmonization, yielding significant improvements in performance [6, 7, 12, 13, 18, 21, 33]. Methods such as DoveNet [8] and BargainNet [6] treat image harmonization as a domain translation problem, aiming to enhance domain consistency between the background and foreground objects. By leveraging advanced deep learning techniques, these models can better capture complex relationships and adapt to various scenarios. Ling *et al.* [23] introduced the Region-aware Adaptive Instance Normalization (RAIN) module, an innovative approach that transfers the statistical properties of background features to normalized foreground features. RAIN has demonstrated promising results, particularly in terms of its ability to harmonize composite images with diverse content. Other methods such as S^2CRNet and CDTNet have integrated color transformation into image harmonization tasks,
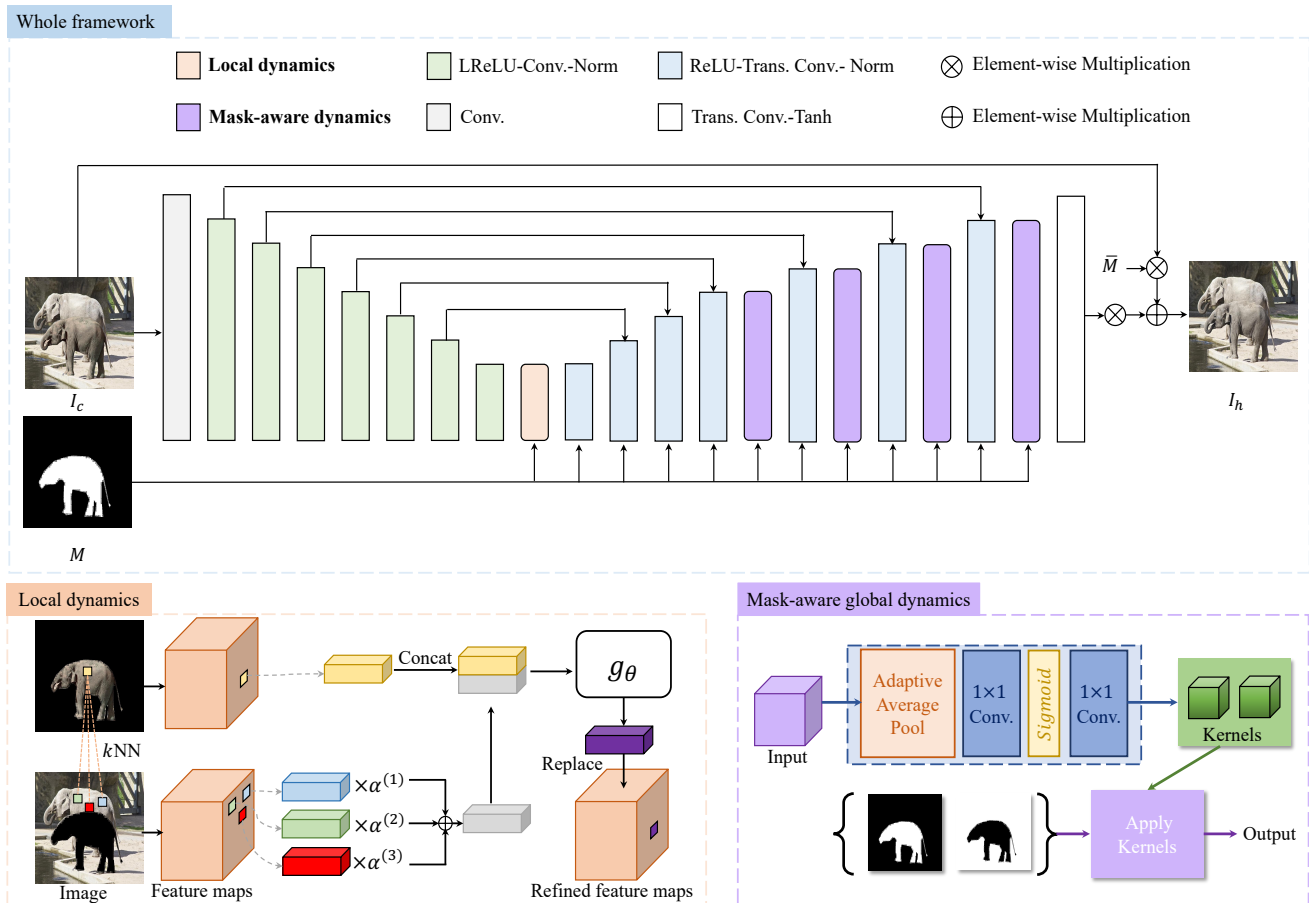
Figure 2: Overview of our proposed hierarchical dynamic image harmonization model. The HDNet consists of an Encoder, a Decoder, a Local Dynamic module and several Mask-aware Global Dynamic modules. Local dynamic module fuses each foreground local representation with its K-nearest neighbor background local representations to achieve local visual consistency. Mask-aware global dynamic module aims to learn the representations of foreground and background regions as well as their correlations from the global view, facilitating global visual consistency for the images much more efficiently.

with their models specifically designed to handle high-resolution problems.

Nevertheless, a common limitation of many existing methods is their reliance on global features for transformation, which can be ineffective for image harmonization tasks that require finer control over local details. In response to this challenge, our proposed model adopts a hierarchical approach that progressively adapts features from a local to a global view. By incorporating both local and global information, our model is better suited to harmonize images with varying levels of complexity and detail, ultimately resulting in more natural and visually coherent composites.

2.2 Style Transfer

Style transfer is a technique that seeks to modify the stylistic appearance of an image based on given style patterns while maintaining the original content structure. This process has been the focus of numerous studies, leading to the development of various methods that achieve style transfer in different ways.

Huang *et al.* [16] proposed Adaptive Instance Normalization (AdaIN), which applies channel-wise mean and variance of style features to align the distribution of content and style images as closely as possible. This method allows for effective style transfer while preserving the underlying content structure. Another approach, Batch-Instance Normalization (Batch-IN)[25], combines the concepts of batch normalization and instance normalization to achieve style transfer. Jing *et al.*[17] introduced dynamic instance normalization, a technique that generates weights using a learnable network that takes the style image as input. This method offers a more flexible and adaptive way to transfer styles. In the Whitening and Coloring Transform (WCT)[19] method, style transfer is achieved by first whitening the content representation and subsequently coloring it with the style representation. This approach allows for a more direct manipulation of the content and style features. The Style Attentional Network (SANet)[26] focuses on efficiently and flexibly integrating local style patterns based on the semantic spatial distribution of the content image. This approach

results in a more coherent and natural-looking style transfer. The recent RainNet method [23] has demonstrated the effectiveness of AdaIN [16] in image harmonization tasks. By employing style transfer techniques, RainNet effectively harmonizes the appearance of composite images while preserving their content structure. This approach highlights the potential of style transfer methods in enhancing image harmonization and underscores the need for further research into the interplay between these two fields.

2.3 Dynamics in Computer Vision

Dynamic networks [4, 10, 14, 20, 34] focus on improving the representation ability of deep models by adapting model structures or parameters during inferences. At their core, dynamic models adaptively apply different weights to parameters or features of the models based on the input data. This adaptability results in increased model capacity and representation ability, enabling the models to better capture complex relationships and patterns in the data. Attention modules [34] are typical examples of dynamic networks, where attention maps are calculated to focus on the important channel or salient region. However, adapting important individually for each pixel may lose the translation invariance of the CNN. To solve this problem, Dynamic Region-Aware Convolution [4] is proposed to assign multiple convolutional filters to different regions separately and share the same filters in each region. Moreover, Deformable Convolution [3, 10, 37] added an offset variable to the position of each sampling point in the convolution kernel, which can realize random sampling nearly points.

3 METHODOLOGY

Our goal is to learn a hierarchical dynamic network for image harmonization. To achieve this goal, we introduce two sub-modules for improving the performance of basic networks, *i.e.*, local dynamic (LD) module and mask-aware global dynamic (MGD) module.

3.1 Overview

Image harmonization aims to adjust the appearance of the foreground object to make it compatible with the background. In a typical image harmonization task, we are given a foreground image I_f and a background image I_b. The foreground mask, denoted by M, indicates the region to be harmonized in the composite image. The composite image I_c can be formulated as $I_c = M \times I_f + (1 - M) \times I$. It is worth noting that the background mask can be represented as $\overline{M} = 1 - M$. Following [30], we only employ the foreground MSE loss as our loss function:

$$\mathcal{L}(I, I_h) = \frac{\sum\limits_{y,x} ||I^{y,x} - I_h{}^{y,x}||}{\text{Max}\{A_{min}, \sum\limits_{y,x} M^{y,x}\}}. \tag{1}$$

A_{min} is a hyperparameter for preventing instability during training and $I^{y,x}$ is the ground truth.

Figure 2 shows the overall framework of our method. Following [6, 7, 16, 18], we employ the U-Net structure as our generator G to harmonize the foreground. The generator G contains an encoder E, a decoder H, a local dynamic module and several mask-aware global dynamic modules. The details of our network structure can be found in supplementary.

3.2 Local Dynamic Module

Considering the significantly variant appearances of different regions of the foreground and the background, recent methods with global feature alignment may not be effective enough in image harmonization tasks. Thus a background location that exhibits greater similarity to the foreground location require more attention, as they play a crucial role in achieving visually coherent harmonization results. To achieve this, we propose a local dynamic module that can adaptive adjust the appearance of each foreground location by matching and fusing related background locations.

For each image, after passed through the encoder, we can obtain its deep features $F \in \mathbb{R}^{C \times H \times W}$ and the corresponding resized foreground mask $M \in \mathbb{R}^{C \times H \times W}$, where C, H, W indicate the number of channels, height, and width of F, respectively. The encoder feature F can be viewed as a set of $H \times W$ C-dimensional local representations. By utilizing the mapping relationship of the mask, these local representations can also be divided into foreground local representations $F_f \in \mathbb{R}^{C \times N_f}$ and background local representations $F_b \in \mathbb{R}^{C \times N_b}$, where N_f is the number of foreground local representations, N_b the number of background local representations and $N_f + N_b = HW$. For each foreground local representations, we aim to find background local representations with similar appearance and semantics and use these background local representations to adjust its appearance. We first calculate similarity map S as below:

$$S^{(i,j)} = \cos(F_f^{(i)}, F_b^{(j)}), \tag{2}$$

$$\cos(F_f^{(i)}, F_b^{(j)}) = \frac{F_f^{(i)\top} F_b^{(j)}}{||F_f^{(i)}|| \cdot ||F_b^{(j)}||}, \tag{3}$$

where $i \in \{1, ..., N_f\}$, $j \in \{1, ..., N_b\}$, $S^{(i,j)}$ is the distance between the i-th local representation of the foreground image and the j-th local representation of background image and $cos(\cdot, \cdot)$ is the cosine similarity. For each foreground local representations, we select its K-nearest neighbors in background, and fuse these local representations to one reference representation $\phi_{ref}^{(i)}$:

$$\phi_{ref}^{(i)} = \sum_{k=1}^{K} \alpha^{(k)} \times F_b^{(k)}, \tag{4}$$

where $\sum_{k=1}^{K} \alpha^{(k)} = 1$, $\alpha^{(k)} > 0$ and $k = 1, ..., K$. The $\alpha^{(k)}$ is obtained by applying the softmax function to the selected K local representations. Then we use these reference local representations to adaptively adjust the corresponding foreground local representations. We concatenate the foreground and reference local representations together and fuse them through an adaptive layer:

$$\phi_{fuse} = g_\theta(\text{Concat}(\phi_{ref}, F_f)), \tag{5}$$

where ϕ_{fuse} is the fused local representations and g_θ indicates the adaptive layer.

3.3 Mask-aware Global Dynamic Module

Recent works [6, 9, 23] show that using attention blocks in the decoder helps improve performance. However, it may not be effective to perform spatial attention on hybrid encoder-decoder features

Model	Param.	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 [33]	41.76MB	92.65	32.28	163.38	29.55	51.85	34.69	82.34	34.62	76.77	33.41
S^2AM [9]	66.70MB	63.40	33.77	143.45	30.03	41.07	35.47	76.61	34.50	59.67	34.35
iS^2AM [30]	66.70MB	21.60	38.28	69.43	33.65	16.15	39.40	40.39	37.87	24.13	38.41
DoveNet [8]	54.76MB	52.32	34.34	133.14	30.21	36.72	35.83	51.95	35.27	52.33	34.76
RainNet [23]	54.75MB	43.35	36.22	110.59	31.64	29.52	37.08	57.40	34.83	40.29	36.12
BargainNet [6]	58.74MB	39.94	35.34	97.32	31.34	24.84	37.03	50.98	35.67	37.82	35.88
Intrinsic [13]	40.86MB	43.02	35.20	105.13	31.34	24.92	37.16	55.53	35.96	38.71	35.90
D-HT [12]	27.00MB	38.53	36.88	74.51	33.13	16.89	38.76	53.01	37.10	30.30	37.55
Harmonizer [18]	21.70MB	21.89	37.64	64.81	33.63	17.34	38.77	33.14	37.56	24.26	37.84
S^2CRNet-SN [21]	**0.95MB**	44.52	35.93	115.46	31.63	28.25	37.65	53.33	36.28	43.20	36.45
S^2CRNet-VGG [21]	15.14MB	34.91	36.42	98.73	32.48	23.22	38.48	51.67	36.81	35.58	37.18
SCS-Co [15]	-	21.01	38.29	**55.83**	34.22	**13.58**	**39.88**	41.75	37.83	**21.33**	38.75
DCCF [35]	-	23.34	37.75	64.77	33.60	17.07	38.66	55.76	37.40	24.65	37.87
CDTNet [7]	-	20.62	38.24	68.61	33.55	16.25	39.15	36.72	37.95	23.75	38.23
SP-IC cycle [1]	66.70MB	**18.17**	**38.91**	68.85	33.88	14.82	39.73	41.32	37.90	22.47	**38.81**
HDNet-lite	**0.65MB**	24.94	39.16	63.55	**34.30**	17.33	39.21	**32.73**	**38.36**	24.99	38.63
HDNet	10.41MB	**13.58**	**41.17**	**47.39**	**35.81**	**11.60**	**41.04**	**31.97**	**38.85**	**16.55**	**40.46**

Table 1: Quantitative comparison across four sub-datasets of iHarmony4 [8]. Top two performance are shown in red and blue. ↑ means the higher the better, and ↓ means the lower the better.

Model	Param.	0% ~5%		5% ~15%		15%~100%		Average	
		MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
Composite	-	28.51	1208.86	119.19	1323.23	577.58	1887.05	172.47	1387.30
DIH [33]	41.76MB	18.92	799.17	64.23	725.86	228.86	768.89	76.77	773.18
S^2AM [9]	66.70MB	13.51	509.41	41.79	454.21	137.12	449.81	48.00	481.79
DoveNet [8]	54.76MB	14.03	591.88	44.90	504.42	152.07	505.82	52.36	549.96
RainNet [23]	54.75M	11.66	550.38	32.05	378.69	117.41	389.80	40.29	469.60
iS^2AM [30]	66.70MB	6.35	288.19	19.69	226.00	71.68	235.30	24.13	260.22
BargainNet [6]	58.74MB	10.55	450.33	32.13	359.49	109.23	353.84	37.82	405.23
Intrinsic [13]	40.86MB	9.97	441.02	31.51	363.61	110.22	354.84	38.71	400.29
S^2CRNet-SN [21]	**0.95M**	8.42	301.97	29.74	336.24	126.56	405.13	43.21	336.99
S^2CRNet-VGG [21]	15.14MB	6.80	**239.94**	25.37	271.70	103.42	333.96	35.58	274.99
SP-IC cycle [1]	66.70MB	**6.08**	276.59	**18.27**	**209.56**	**66.44**	**216.37**	22.47	**245.75**
HDNet-lite	**0.65MB**	6.45	289.71	18.70	217.66	76.56	243.58	24.99	260.65
HDNet	10.41MB	**4.35**	**199.56**	**12.82**	**150.45**	**49.18**	**159.82**	**16.55**	**179.49**

Table 2: We measure the error of different methods in foreground ratio range based on the whole test set. fMSE indicates the mean square error of the foreground region.

since pixel-level adaptation is unsuitable for low-level texture features.

To learn adaptive representations for harmonious and inharmonious regions, we propose the mask-aware global dynamic module to predict the adaptive convolutional kernels with the guidance of the mask. As shown in Figure 2, mask-aware global dynamic module are incorporated into our networks to better integrate the

local information for modeling the visual coherence. Unlike DR-conv [4] focusing on the local information, which is unreliable in the harmonious and inharmonious regions, we learn different kernels according to the foreground mask. For efficiency, different groups of filters for foreground and background are applied for the whole input to get the dynamic features. Then the dynamic features are multiplied by the mask. Finally, a summation is applied
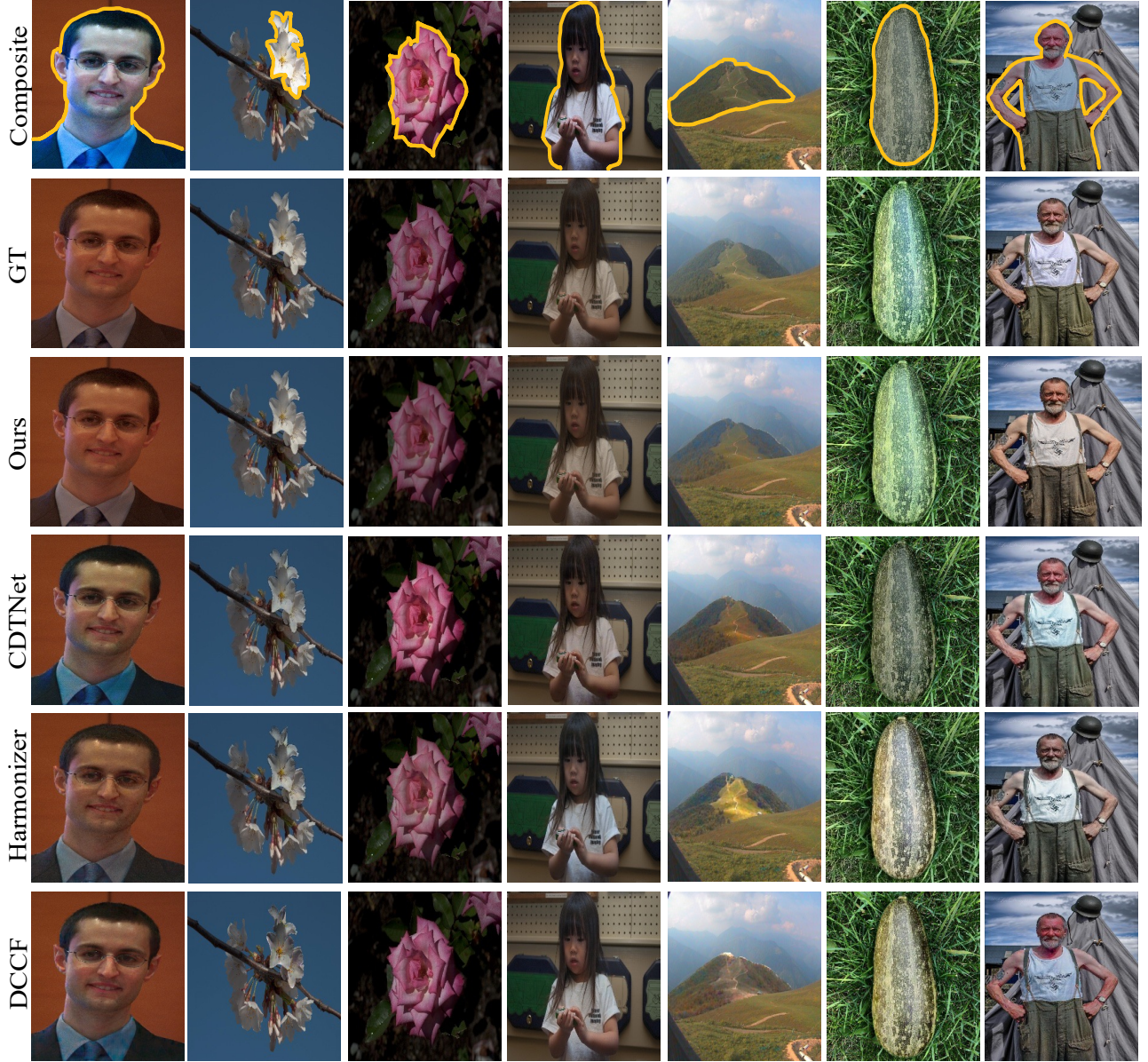
Figure 3: Qualitative comparison on samples from the testing dataset of iHarmony4. The yellow border lines indicate the foreground.

to obtain the final results for the MGD:

$$F'_m = (F_m \odot W_f) \otimes M + (F_m \odot W_b) \otimes \overline{M}, \qquad (6)$$

where $\odot$ denotes covolution operation, W_f and W_b are the filters.

Why would our model work? Since LD integrates foreground local representations with the K-nearest neighbors of background local representations, effectuating adaptive transfer of the background appearance to the foreground. Concurrently, MGD applies different kernels on the foreground and background regions, each region can be regarded as being assigned an individual decoder to learn the harmonization mapping, but without introducing extra computational cost, since all regions share the same encoder for feature extraction.

4 EXPERIMENTS

In this section, we first introduce the datasets, metrics, and implementation details for our experiments. We then compare HDNet with existing image harmonization methods. We further conduct ablation experiments to evaluate the effectiveness of individual modules in HDNet. Finally, we demonstrate the advantages of HDNet in real-world image harmonization applications.
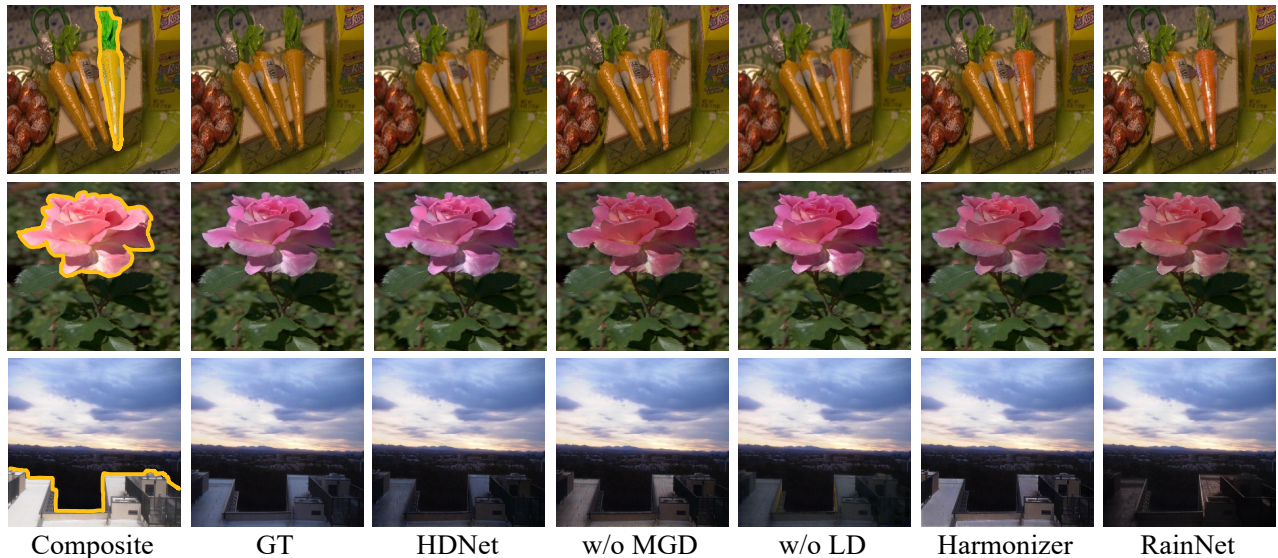
Composite   GT   HDNet   w/o MGD   w/o LD   Harmonizer   RainNet

Figure 4: Ablation study on samples from the testing dataset of iHarmony4.

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
Base	24.33	38.83	65.77	34.05	17.51	39.27	34.08	38.15	25.20	38.53
+LD	14.79	**40.96**	51.51	35.57	12.44	40.82	32.79	38.45	17.64	40.23
+SA	15.89	40.46	55.46	35.33	13.08	40.38	32.30	39.00	19.19	39.81
+MGD	16.37	40.81	55.56	35.36	12.82	40.75	32.44	**38.82**	18.99	40.13
HDNet*	**14.39**	41.03	**48.46**	**35.79**	**11.94**	**40.91**	**30.69**	38.80	**17.08**	**40.34**
HDNet	**13.58**	**41.17**	**47.39**	**35.81**	**11.60**	**41.04**	**31.97**	**38.85**	**16.55**	**40.46**

Table 3: Ablation study on iHarmony4 [8]. HDNet* indicates that we use the learned mask instead of the original mask provided by datasets. SA indicates the spatial attention module used in RainNet [23] and SCS-Co [15]

Model	Parms.	Time(s)	PSNR↑
S^2AM [9]	66.70M	0.25	34.35
DoveNet [8]	54.76M	0.05	34.76
BargainNet [6]	58.74M	0.21	35.88
Intrinsic [13]	40.86M	1.17	35.90
S^2CRNet-SN [21]	**0.95M**	**0.03**	**36.45**
HDNet-lite	**0.65M**	**0.04**	**38.63**

Table 4: Average processing time on the CPU.

4.1 Experiment Setting

Datasets. Following the recent works [6, 8, 23], we conduct image harmonization tasks on iHarmony4 benchmark [8]. iHarmony4 includes 73,146 image pairs for image harmonization and contains four subsets: HAdobe5k, HFlickr, HCOCO, and Hday2night. Each sample in iHarmony4 consists of a natural image, a foreground mask, and a composite image (with the foreground generated by

GAN [11]). We follow the same partition settings of iHarmony4 as DoveNet [8]. Note that we conduct high-resolution (i.e., 1024×1024 and 2048 × 2048) experiments on HAdobe5k since only HAdobe5k contains high-resolution images.

Implementation Details. HDNet is trained from scratch by Adam optimizer with $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The batch size is set to 12 and we train our HDNet for 120 epochs. The initial learning rate is set to 0.001. The initial learning rate is multiple by 0.1 in the 100-th and 110-th epochs. All images are resized to 256×256, batch size set to 12, and no data augmentations are adopted. We use PyTorch [27] to implement our models with Nvidia Tesla A100 GPUs.

Evaluation. During the test phase, we use Mean Square Error (MSE), foreground MSE (fMSE), Structural SIMilarity (SSIM), and Peak Signal-to-Noise Ratio (PSNR) to evaluate the performance. To illustrate performance, we qualitatively compare our method with numerous state-of-the-art methods, including DIH [33], S^2AM [9], iS^{2}AM [30], DoveNet [8], RainNet [23], Bargainnet [6], Intrinsic [13], D-HT [12], CDTNet [7], Harmonizer [18], DCCF [35], S^2CRNet [21], SCS-CO [15], SP-IC cycle [1] and INR[5].

4.2 Comparison with Other Methods

Performances on different sub-datasets. Table 1 lists the quantitative results of previous state-of-the-art methods and our method. From Table 1, we can observe that our method outperforms all of them across all sub-datasets and all metrics. Compared to the most recent TMM'23 method SP-IC cycle [1] on iHarmony4 dataset, our HDNet brings 2.92 improvement in terms of MSE, and 1.65 dB improvement in terms of PSNR.

Moreover, to make our model practical, that is, it can be used on edge devices (e.g., mobile phones), we propose HDNet-lite. HDNet-lite is obtained by reducing the number of HDNet channels to 1/4. Compared to the method with equivalent performance, our HDNet-lite has fewer parameters. For example, compared to RainNet and BarginNet, HDNet-lite only uses 1.2% of the parameters to achieve better performance in the PSNR metric, demonstrating the effectiveness of the proposed network.

Influence of foreground ratios. Following [23], we examine the influence of different foreground ratios on the harmonization models, i.e., 0% to 5%, 5% to 15%, 15% to 100%, and overall results. The results of all previous methods and our HDNet are given in Table 2. It can be observed that our HDNet achieves the best performance among all approaches. HDNet works well at various foreground scales, thanks to its combination of hierarchical dynamics.

Qualitative comparisons. We take a closer look at model performance and provide qualitative comparisons with the previous competing methods. From the sample results in Figure 3, it can be easily observed that our approach integrates the foreground objects into the background image, achieving much better visual consistency than other methods. Our HDNet can achieve these photorealistic results because our HDNet adaptively adjusts the feature of foreground and background by hierarchical dynamics learning.

4.3 Ablation Study

Visual comparison. To further illustrate the effectiveness of our hierarchical dynamics, we show some output results of ablation experiments in Figure 4. It can be found that compared with the distortion results produced by the baseline, after adding our proposed dynamics, the color and lighting of the output results are close to the real images. Each dynamics contribute to the final result because they conduct dynamic learning at different feature levels.

Effectiveness of local dynamic module. Our local dynamic module adjusts the appearance of each foreground local representation according to the K-nearest background local representations. In Table 3, we can see that adding LD to the baseline brings 1.7 dB and 7.56 average performance improvement in terms of PSNR and MSE. Moreover, if we remove LD from HDNet, the PSNR will decrease by 0.23 dB and MSE will decrease by 2.44.

Effectiveness of mask-aware global dynamic module. Our mask-aware global dynamic module integrates the local information to model visual coherence. Adding MGD to the model will bring significant improvement, proving that it is not effective enough to perform spatial attention [15, 23] on hybrid encoder-decoder features since pixel-level adaptation is unsuitable for such low-level texture features. Moreover, if we use the learned mask to replace

Model	Resolution	PSNR↑	MSE↓	fMSE↓	SSIM↑
Composite		352.05	28.10	2122.37	0.9642
DoveNet [8]		34.81	51.00	312.88	0.9729
S²AM [9]		35.68	47.01	262.39	0.9784
Intrinsic [13]	1024×1024	34.69	56.34	417.33	0.9471
RainNet [23]		36.61	42.56	305.17	0.9844
CDTNet [7]		**38.77**	**21.24**	**152.13**	0.9868
INR [5]		38.38	22.68	187.97	**0.9886**
HDNet		**41.56**	**13.24**	**102.53**	**0.9931**
CDTNet [7]		37.66	29.02	198.85	0.9845
INR [5]	2048×2048	**38.35**	**24.08**	**192.20**	**0.9886**
HDNet		**41.29**	**18.35**	**147.25**	**0.9911**

Table 5: High-resolution experiments on HAdobe5K.

the original mask, the performance will decline, indicating that the learned mask is unreliable.

Harmonization performance on CPU. Our HDNet shows relatively fast processing speed on CPU devices, which enables our method to run on the device side without any cloud computation. To this end, we compare the proposed HDNet with other baseline methods [6, 8, 9, 12, 21] in harmonizing under the same experimental environment (Intel Xeon Platinum 8369B CPU on Ubuntu 18.04). The evaluations are conducted on the 50 images in the HAdobe5k sub-dataset and we present the average processing time in Table 4. The experimental results show that our method has the second fastest inference speed when inference on CPU but the performance of our model is much better than other methods.

High-resolution results. Following [7], we conduct high-resolution image harmonization experiments. As shown in Table 5, we can see that our method outperforms all of them across all metrics. Compared with the most recent method INR [5], under 1024×1024 resolution setting, our method achieves a huge average performance gain of 9.44 in MSE, 85.44 in fMSE, 0.0045 in SSIM, and 3.18 in PSNR.

Influence of neighbors. In the local dynamic module, we need to find the K-nearest neighbors in the background image for each local representation of the foreground image. Next, we fuses every foreground local representations with its K-nearest background local representations. How to choose a suitable hyperparameter k is thus a key. For this purpose, we perform a low-resolution harmonization task (i.e., 256×256) by varying the value of $K \in \{1, 3, 5, 7, 9\}$, and show the results in Figure 5. It can be observed that the performance is best when K is equal to 1. This may be attributed to the provision of potentially negative information by an excessive number of local representations.

5 CONCLUSION

This paper proposes a hierarchical dynamic network (HDNet) from local to global that gradually builds local dynamic module and mask-aware global dynamic module. Local dynamic module fuses each foreground local representation with its K-nearest neighbor background local representations to achieve local visual consistency. Mask-aware global dynamic module aims to learn representations
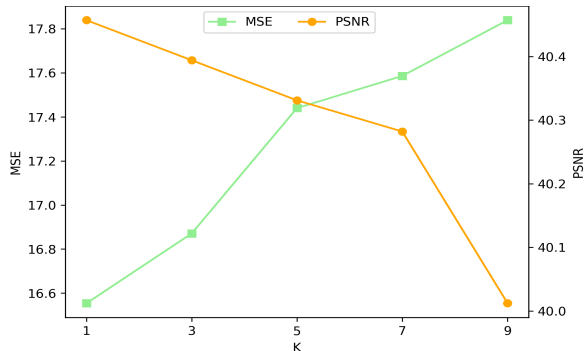
Figure 5: Influence of neighbors. The PSNR score decreases and MSE increases with larger K.

of foreground and background regions as well as their correlations from the global view, facilitating global visual consistency for the images much more efficiently. Our method achieves state-of-the-art performances on the benchmark dataset iHarmony4 and our light-weight version model HDNet-lite achieves state-of-the-art results compared to other methods while only has 0.65MB parameters. Our limitation mainly lies in the dependency on mask. When a disharmonious image does not provide a mask, the performance of using a learning mask is not good. Future investigation into these issues should be required.

REFERENCES

[1] Xun Cai, Qingjie Shi, Yanbo Gao, Shuai Li, Wei Hua, and Tian Xie. 2023. A Structure-Preserving and Illumination-Consistent Cycle Framework for Image Harmonization. *IEEE Trans. Multim.* (2023).

[2] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. 2021. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*. 324–333.

[3] Feng Chen, Fei Wu, Jing Xu, Guangwei Gao, Qi Ge, and Xiao-Yuan Jing. 2021. Adaptive deformable convolutional network. *Neurocomputing* 453 (2021), 853–864.

[4] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. 2021. Dynamic region-aware convolution. In *CVPR*. 8064–8073.

[5] Jianqi Chen, Yilan Zhang, Zhengxia Zou, Keyan Chen, and Zhenwei Shi. 2023. Dense Pixel-to-Pixel Harmonization via Continuous Image Representation. *arXiv preprint arXiv:2303.01681* (2023).

[6] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. 2021. Bargainnet: Background-Guided Domain Translation for Image Harmonization. In *ICME*. 1–6.

[7] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. 2022. High-Resolution Image Harmonization via Collaborative Dual Transformations. In *CVPR*. 18470–18479.

[8] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. 2020. DoveNet: Deep Image Harmonization via Domain Verification. In *CVPR*. 8391–8400.

[9] Xiaodong Cun and Chi-Man Pun. 2020. Improving the Harmony of the Composite Image by Spatial-Separated Attention Module. *IEEE Trans. Image Process.* 29 (2020), 4759–4771.

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. 2017. Deformable convolutional networks. In *ICCV*. 764–773.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *NeurIPS*. 2672–2680.

[12] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. 2021. Image Harmonization with Transformer. In *ICCV*. 14850–14859.

[13] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. 2021. Intrinsic Image Harmonization. In *CVPR*. 16367–16376.

[14] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. 2021. Dynamic neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* (2021).

[15] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. 2022. Scs-co: Self-consistent style contrastive learning for image harmonization. In *CVPR*. 19710–19719.

[16] Xun Huang and Serge J. Belongie. 2017. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In *ICCV*. 1510–1519.

[17] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. 2020. Dynamic Instance Normalization for Arbitrary Style Transfer. In *AAAI*. 4369–4376.

[18] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. 2022. Harmonizer: Learning to Perform White-Box Image and Video Harmonization. In *ECCV*. 690–706.

[19] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. 2017. Universal style transfer via feature transforms. *NeurIPS* 30 (2017).

[20] Yaohui Li, Yuzhe Yang, Huaxiong Li, Haoxing Chen, Liwu Xu, Leida Li, Yaqian Li, and Yandong Guo. 2022. Transductive Aesthetic Preference Propagation for Personalized Image Aesthetics Assessment. In *ACM MM*.

[21] Jingtang Liang, Xiaodong Cun, and Chi-Man Pun. 2022. Spatial-Separated Curve Rendering Network for Efficient and High-Resolution Image Harmonization. In *ECCV*.

[22] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. 2021. Editgan: High-precision semantic image editing. *NeurIPS* 34 (2021), 16331–16345.

[23] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. 2021. Region-Aware Adaptive Instance Normalization for Image Harmonization. In *CVPR*. 9361–9370.

[24] J. Matías Di Martino, Gabriele Facciolo, and Enric Meinhardt-Llopis. 2016. Poisson Image Editing. *Image Process. Line* 6 (2016), 300–325.

[25] Hyeonseob Nam and Hyo-Eun Kim. 2018. Batch-Instance Normalization for Adaptively Style-Invariant Neural Networks. In *NeurIPS*. 2563–2572.

[26] Dae Young Park and Kwang Hee Lee. 2019. Arbitrary style transfer with style-attentional networks. In *CVPR*. 5880–5888.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS* 32 (2019).

[28] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. 2001. Color Transfer between Images. *IEEE Computer Graphics and Applications* 21, 5 (2001), 34–41.

[29] Jing Shi, Ning Xu, Haitian Zheng, Alex Smith, Jiebo Luo, and Chenliang Xu. 2022. SpaceEdit: Learning a Unified Editing Space for Open-Domain Image Color Editing. In *CVPR*. 19730–19739.

[30] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. 2021. Foreground-aware semantic representations for image harmonization. In *WACV*. 1620–1629.

[31] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. 2010. Multi-scale image harmonization. *ACM Trans. Graph.* 29, 4 (2010), 1–10.

[32] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. 2022. Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In *AAAI*, Vol. 36. 2352–2360.

[33] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep Image Harmonization. In *CVPR*. 2799–2807.

[34] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *ECCV*. 3–19.

[35] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. 2022. DCCF: Deep Comprehensible Color Filter Learning Framework for High-Resolution Image Harmonization. In *ECCV*.

[36] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly E. Rushmeier. 2012. Understanding and improving the realism of image composites. *ACM Trans. Graph.* 31, 4 (2012), 84:1–84:10.

[37] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. 2019. Deformable ConvNets V2: More Deformable, Better Results. In *CVPR*. 9308–9316.