Hierarchical Dynamic Image Harmonization

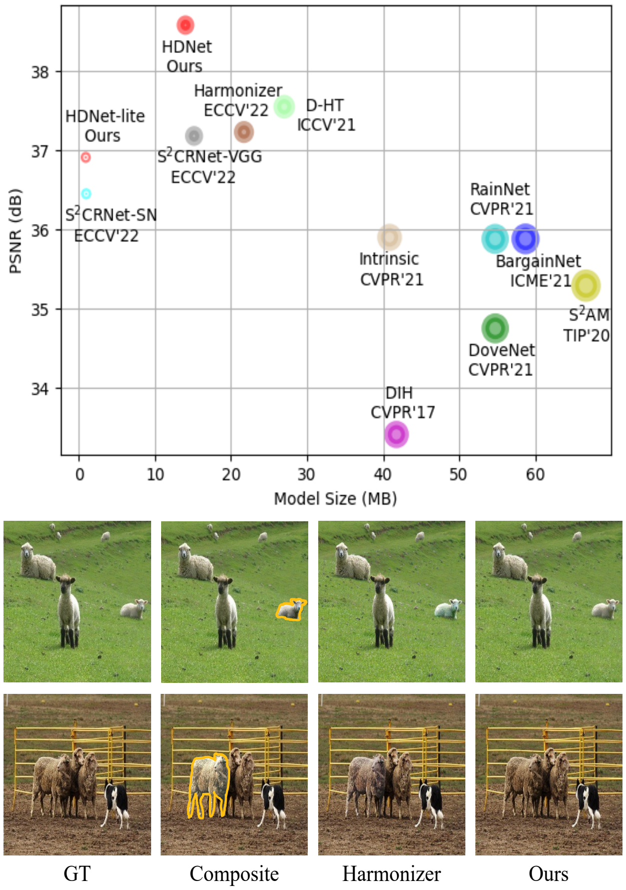
Haoxing Chen[1,2], Zhangxuan Gu[2], Yaohui Li[1], Jun Lan[2]
Changhua Meng[2], Weiqiang Wang[2], Huaxiong Li[1]
[1]Department of Control Science and Intelligence Engineering, Nanjing University
[2]Tiansuan Lab, Ant Group
{haoxingchen,yaohuili}@smail.nju.edu.cn, huaxiongli@nju.edu.cn
{guzhuangxuan.gzx,yelan.lj,changhua.mch,weiqiang.wwq}@antgroup.com

Abstract

Image harmonization is a critical task in computer vision, which aims to adjust the fore-ground to make it compatible with the back-ground. Recent works mainly focus on using global transformation (i.e., normalization and color curve rendering) to achieve visual consistency. However, these model ignore local consistency and their model size limit their harmonization ability on edge devices. Inspired by the dynamic deep networks that adapt the model structures or parameters conditioned on the inputs, we propose a hierarchical dynamic network (HDNet) for efficient image harmonization to adapt the model parameters and features from local to global view for better feature transformation. Specifically, local dynamics (LD) and mask-aware global dynamics (MGD) are applied. LD enables features of different channels and positions to change adaptively and improve the representation ability of geometric transformation through structural information learning. MGD learns the representations of fore- and back-ground regions and correlations to global harmonization. Experiments show that the proposed HDNet reduces more than 80% parameters compared with previous methods but still achieves the state-of-the-art performance on the popular iHarmony4 dataset. Our code is avaliable in https://github.com/chenhaoxing/HDNet.

1. Introduction

Combining image patch of a different image into a realistic image is a fundamental technique in the computer vision community, i.e., image editing [20, 27] and scene completion [1, 30]. However, the composite image inevitably suffers from the inharmony problem, since the fore- and background appearance will be distinct due to different imaging conditions (e.g., rainy and sunny, morning and dusk). Thus, image harmonization, which aims at achieving visual consistency within the composite image, is an important and challenging task.



GT Composite Harmonizer Ours

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.

Traditional methods focus on better transferring handcrafted low-level appearance statistics, such as color statistics [26, 35], and gradient information [22]. However, they could not handle the complex situation where the source image has a large appearance or semantic gap with the target.

With the advances in deep learning, more deep neural network-based methods were proposed [4–6, 15]. Most methods use complex network structures or training strategies [4, 6]. In contrast, color transformation and normalization based have received extensive attention due to their simplicity and flexibility [5, 18, 21].

Color transformation based methods aim to learn an RGB-to-RGB transformation. For example, Collaborative Dual Transformation (CDTNet) [5] combine low-resolution generator for pixel-to-pixel transformation, a lookup tables (LUTs) for RGB-to-RGB transformation, and a refinement module to take advantage of both.Spatial-Separated Curve Rendering Network (S^2CRNet) [18] design a curve rendering module (CRM), which learns and combines the spatial-specific knowledge using linear layers to generate the parameters of the piecewise curve mapping in the fore-ground region.

Normalization based methods regard image harmonization as a back-ground-to-fore-ground style transfer problem. Inspired by AdaIN [13], Ling *et al.* [21] regard image harmonization as a back-ground to fore-ground style transfer problem and proposed region-aware adaptive instance normalization (RAIN) which captures the style statistics information from the back-ground features and applies it to the fore-ground. However, as shown in Figure 1, unwanted patterns still exist or even is very severe in some cases.

However, most methods use global features for transformation and are ineffective for image harmonization. These methods ignore that, for a real image, the appearance of different regions can vary significantly. 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.

Inspired by the dynamic deep networks [3, 8, 19, 33] that adapt the model structures or parameters conditioned on the inputs, we solve the above problems by proposing an efficient dynamic image harmonization network, which hierarchically adapts the parameters and features of the model by two dynamics, *i.e.*, local dynamics and mask-aware global dynamics from local to global view. Local dynamics enable channel-wise and spatial-wise variations for representations and local dynamics model the geometric transformations and augment sampling locations for describing the structures. Mask-aware dynamics aim to apply different filters to foreg- and back-ground regions adaptively and keep translation invariance in each region. 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, and meanwhile its performance is 0.42 dB higher than the second-best method.

The main contributions can be summarized as follows:

- We propose a novel hierarchical dynamic image harmonization network, which hierarchically adaptively tunes the parameters and features of the model by local dynamics and mask-aware global dynamics from local to the global view.
- A local dynamic module is designed to obtain channel- and spatial-wise representation variations, and enable modeling geometric transformations to better describe the structures.
- A mask-aware global dynamic module is designed to learn the representations of fore- and back-ground regions as well as their correlations to the global harmonization, facilitating local visual consistency for the images much more efficiently.
- 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 aim at improving composite images via low-level appearance features, such as performing color transformations to match visual appearance [26, 35] or manipulating multi-scale transformation and statistical analysis [29].

Recently, more deep learning-based methods have been proposed with notable successes [4, 5, 10, 11, 15, 18, 31]. DoveNet [6] and BargainNet [4] regard image harmonization as a domain translation problem and focus on improving domain consistency between back- and fore-ground. Ling *et al.* [21] proposed Region-aware Adaptive Instance Normalization (RAIN) module, which transfers the statistics from the back-ground features to the normalized fore-ground features. RAIN has achieved promising results. S^2CRNet and CDTNet introduce color transformation into image harmonization tasks, and their model can support high-resolution problems.

However, most methods use global features for transformation and are ineffective for image harmonization. In contrast, our model hierarchically adapts the parameters and features of the model.

2.2. Style Transfer

Style transfer aims at changing the image style according to given style patterns while preserving content structure.

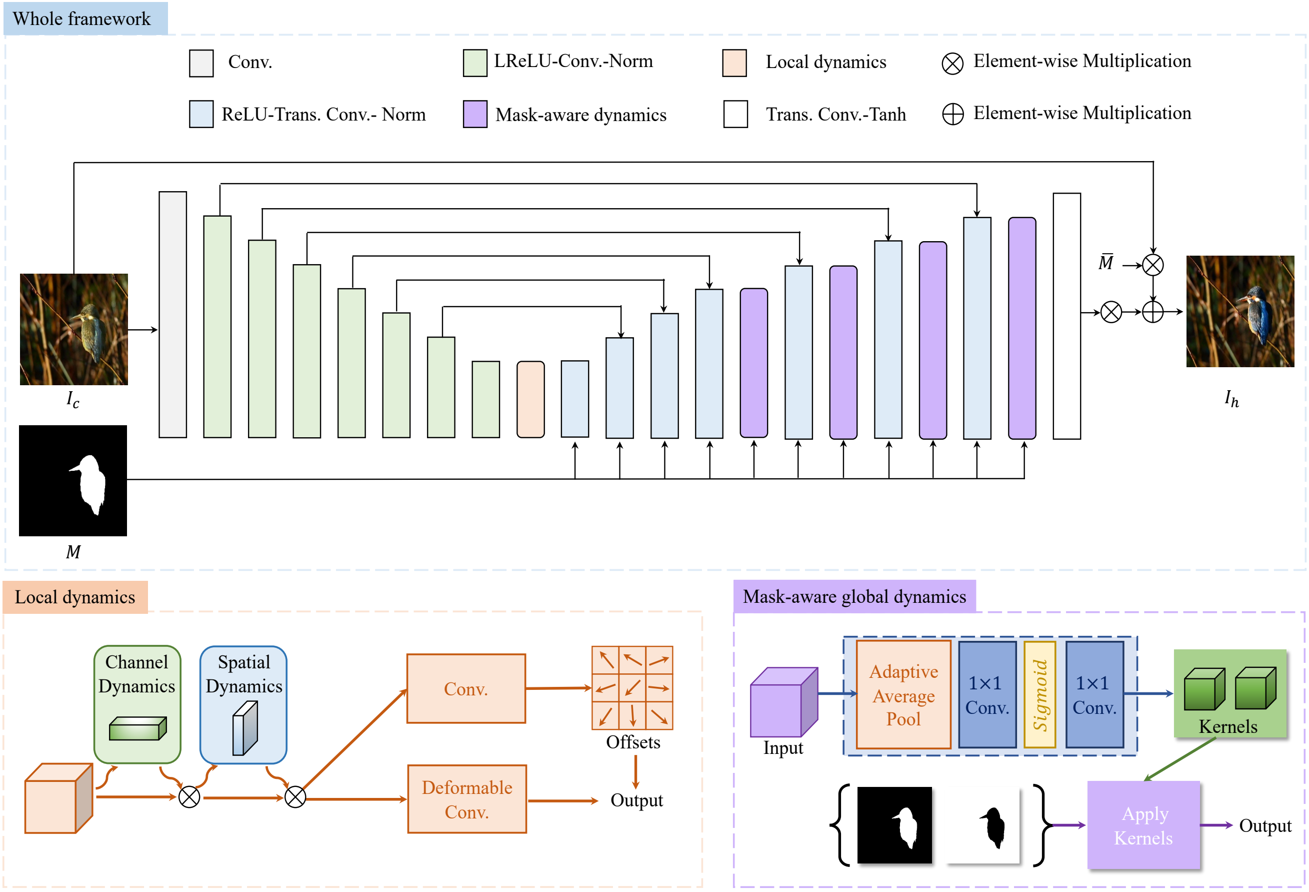


Figure 2. Overview of our proposed hierarchical dynamic image harmonization model.

Huang *et al.* [13] proposed AdaIN that applies channel-wise mean and variance of style feature to make the distribution of content image and style image as close as possible. Batch-IN [23] combines batch normalization and instance normalization. Jing *et al.* [14] proposed dynamic instance normalization that generated weights by a learnable network taking the style image as input. WCT [16] transfer the style by whitening the content representation and then coloring it with style representation. SANet [24] efficiently and flexibly integrates the local style patterns according to the semantic spatial distribution of the content image.

Recent method RainNet [21] also validates that style transfer method AdaIN [13] is very effective in image harmonization.

2.3. Dynamics in Computer Vision

Dynamic networks [3, 8, 12, 17, 33] focus on improving the representation ability of the deep models, which adapt the model structures or parameters during inferences. Essentially, dynamic models adaptively apply different weights on parameters or features of the models conditioned on inputs, which increases the model capacity and representation ability. Attention modules [33] 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 [3] is proposed to assign multiple convolutional filters to different regions separately and share the same filters in each region. Moreover, Deformable Convolution [2,8,36] 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 dynamics (LD) and mask-aware global dynamics (MGD).

3.1. Problem Formulation

Image harmonization aims to adjust the appearance of the fore-ground object to make it compatible with the back-ground. An image harmonization task consists of fore-ground image $I_f \in \mathbb{R}^{3\times H\times W}$ and back-ground image $I_b \in \mathbb{R}^{3\times H\times W}$. The fore-ground mask is denoted by $M \in \mathbb{R}^{1\times H\times W}$, which indicates the region to be harmonized in the composite image $I_c = M \times I_f + (1-M) \times I \in \mathbb{R}^{3\times H\times W}$. Note that the back-ground mask can be denoted as $\overline{M} = 1 - M$. Our goal is to learn a harmonization neural network G, whose output is the harmonized image as $I_h = G(I_c, M)$. Following [28], we only employ the fore-ground MSE loss as our loss function:

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

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

3.2. Hierarchical Dynamic Network (HDNet)

Real-world composite images are always diverse but have local coherence. A hierarchical dynamic network (HDNet) that enables feature dynamics from both local and global views is proposed, gradually building local dynamics and mask-aware global dynamics. As shown in Figure 2, following [4, 5, 13, 15], we employ the U-Net structure as our Generator G to harmonize the fore-ground.

3.2.1 Local Dynamics

Considering the significantly variant appearances of different regions of the fore- and the back-ground, directly applying the global transformation [4, 5, 21]is not effective enough. We argue that the back-ground areas that are feature-similar to the fore-ground need more attention. Inspired by [32, 33], we propose an efficient channel-spatial attention module, which can adapt the feature map channel-wise and spatial-wise which generate channel- and spatial-aware dynamics. Given an input feature F_c, channel-aware dynamics can be calculated by:

$$\mathcal{M}_c = \sigma(\mathrm{Conv1D}_k(\mathrm{GAP}(F_c))), \tag{2}$$

$$F'_c = \mathcal{M}_c \otimes F_c, \tag{3}$$

where $\otimes$ denotes the element-wise multiplication, σ indicates sigmoid function, Conv1D stands for the 1D convolution with kernel size k, GAP is the global-average pooling operation and F'_c is the output feature through channel-aware dynamics.

For spatial-aware dynamics, the adaptation process can be described by:

$$\mathcal{M}_{sp} = \sigma(\mathrm{Conv}(\mathrm{ReLU}(\mathrm{Conv}(F_{sp})))), \tag{4}$$

$$F'_{sp} = \mathcal{M}_{sp} \otimes F_{sp}, \tag{5}$$

where F_{sp} is the input for the spatial-aware component which equals F'_c. The channel-spatial attention module enables features in different channels and different locations varied adaptively, which could benefit visual consistency learning at such a fine-grained feature level.

The channel-spatial attention module enhances local consistency but fails to preserve translational invariance. Varied features from local regions are not good enough to model large irregular variations, especially for the inharmonious images. To better capture the features of local structures, we use adaptively deformable convolutions after the channel-spatial attention module to take advantage of the local consistency of structures and improve the representation ability of geometric transformations. In this case, a learnable offset replaces the regular sampling grid of the convolution. Given a convolutional kernel of K sampling locations, the calculation of the adaptive deformable convolution can be formulated as:

$$F'_s(p) = \sum_{k=1}^{K} w_k \cdot F_s(p + s_k \cdot p_k + \triangle p_k) \cdot ((1 - s_k) \cdot \triangle m_k), \tag{6}$$

where w_k, p_k, $\triangle p_k$, and $\triangle m_k$ refer to wights, handpicked offset, learnable offset and modulation scalar for the k-th location; s_k denotes adaptive dilation factor that contains the general distance information of sampling locations; $F_s(p)$ and $F'_s(p)$ denote the feature representations at location p from the input feature F_s and from calculated structure-aware feature F'_s.

The spatial sampling locations are largely augmented from the learning of the harmonization task which can better represent the features of the irregular structures. After the adaptive deformable convolution component, the focus of the features is obviously enlarged.

3.2.2 Mask-aware Global Dynamics

Recent works [4, 7, 21] 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 since pixel-level adaptation is unsuitable for low-level texture features. Moreover, the structure-aware dynamics mainly focus on local structures as the irregular sampling offsets of adaptive deformable convolutions are still limited to local structure areas. As shown in Figure 2, mask-aware global dynamics are incorporated into our networks to better integrate the local information for modeling the local coherence.

To learn adaptive representations for harmonious- and inharmonious regions, we propose the mask-aware global dynamics module to predict the adaptive convolutional kernels with the guidance of the mask. Unlike DRconv [3]

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 [31]	41.76M	92.65	32.28	163.38	29.55	51.85	34.69	82.34	34.62	76.77	33.41
S^2AM [7]	66.70M	63.40	33.77	143.45	30.03	41.07	35.47	76.61	34.50	59.67	34.35
DoveNet [6]	54.76M	52.32	34.34	133.14	30.21	36.72	35.83	51.95	35.27	52.33	34.76
RainNet [21]	54.75M	43.35	36.22	110.59	31.64	29.52	37.08	57.40	34.83	40.29	36.12
BargainNet [4]	58.74M	39.94	35.34	97.32	31.34	24.84	37.03	50.98	35.67	37.82	35.88
Intrinsic [11]	40.86M	43.02	35.20	105.13	31.34	24.92	37.16	55.53	35.96	38.71	35.90
D-HT [10]	27.00M	38.53	36.88	74.51	33.13	16.89	38.76	53.01	37.10	30.30	37.55
Harmonizer [15]	21.70M	21.89	37.64	64.81	**33.63**	17.34	38.77	**33.14**	37.56	24.26	37.84
S^2CRNet-SN [18]	**0.95M**	44.52	35.93	115.46	31.63	28.25	37.65	53.33	36.28	43.20	36.45
S^2CRNet-VGG [18]	15.14M	34.91	36.42	98.73	32.48	23.22	38.48	51.67	36.81	35.58	37.18
DCCF [34]	-	23.34	37.75	**64.77**	33.60	17.07	38.66	55.76	37.40	24.65	37.87
CDTNet [5]	-	**20.62**	**38.24**	68.61	33.55	**16.25**	**39.15**	36.72	**37.95**	**23.75**	**38.23**
HDNet-lite	**0.89M**	35.33	36.45	82.87	32.10	23.75	38.08	59.58	36.41	34.39	36.91
HDNet	14.01M	**22.67**	**38.56**	**63.85**	**33.96**	**15.59**	**39.49**	**35.92**	**38.11**	**23.42**	**38.58**

Table 1. Quantitative comparison across four sub-datasets of iHarmony4 [6]. 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 [31]	41.76M	18.92	799.17	64.23	725.86	228.86	768.89	76.77	773.18
S^2AM [7]	66.70M	13.51	509.41	41.79	454.21	137.12	449.81	48.00	481.79
DoveNet [6]	54.76M	14.03	591.88	44.90	504.42	152.07	505.82	52.36	549.96
RainNet [21]	54.75M	11.66	550.38	32.05	378.69	117.41	389.80	40.29	469.60
BargainNet [4]	58.74M	10.55	450.33	32.13	359.49	109.23	353.84	37.82	405.23
Intrinsic [11]	40.86M	9.97	441.02	31.51	363.61	110.22	354.84	38.71	400.29
S^2CRNet-SN [18]	**0.95M**	8.42	301.97	29.74	336.24	126.56	405.13	43.21	336.99
S^2CRNet-VGG [18]	15.14M	**6.80**	**239.94**	**25.37**	**271.70**	103.42	333.96	35.58	**274.99**
HDNet-lite	**0.89M**	9.42	431.92	28.77	331.54	**100.32**	**325.68**	**34.39**	382.26
HDNet	14.01M	**5.95**	**230.75**	**20.32**	**265.31**	**68.95**	**318.15**	**23.42**	**258.8**

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

focusing on the local information, which is unreliable in the harmonious- and inharmonious regions, we learn different kernels according to the fore-ground mask. For efficiency, different groups of filters for fore-ground and back-ground 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 to obtain the final results for the MGD:

$$F'_m = (F_m \odot W_f) \otimes M + (F_m \odot W_b) \otimes \overline{M}, \quad (7)$$

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

Why would our model work? Since LD enhances the local interaction between fore- and back-ground through channel-spatial attention and structure information, and MGD applies different kernels on the fore- and back-ground 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 im-

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	25.86	36.68	104.51	32.85	18.60	38.45	47.01	36.46	30.84	37.27
+LD	29.26	38.26	68.55	33.68	**16.00**	39.32	40.65	37.52	26.19	38.35
+MGD	26.69	37.97	71.39	33.64	16.46	39.31	37.97	37.71	25.99	38.25
HDNet*	**24.32**	**38.45**	**66.32**	**33.75**	16.14	**39.43**	**36.65**	**37.85**	**25.42**	**38.45**
HDNet	**22.67**	**38.56**	**63.85**	**33.96**	**15.59**	**39.49**	**35.92**	**38.11**	**23.42**	**38.58**

Table 3. Ablation study on iHarmony4 [6]. HDNet* indicates that we use the learned mask instead of the original mask provided by datasets.

Model	PSNR↑	MSE↓	fMSE↓	SSIM↑
Composite	352.05	28.10	2122.37	0.9642
DoveNet [6]	34.81	51.00	312.88	0.9729
S^2AM [7]	35.68	47.01	262.39	0.9784
Intrinsic [11]	34.69	56.34	417.33	0.9471
RainNet [21]	36.61	42.56	305.17	0.9844
CDTNet [5]	**38.77**	**21.24**	**152.13**	**0.9868**
HDNet	**41.26**	**15.42**	**103.08**	**0.9899**

Table 4. High-resolution experiments on HAdobe5K.

Model	Parms.	Time(s)	PSNR↑
S^2AM [7]	66.70M	0.25	34.35
DoveNet [6]	54.76M	0.05	34.76
BargainNet [4]	58.74M	0.21	35.88
Intrinsic [11]	40.86M	1.17	35.90
S^2CRNet-SN [18]	**0.95M**	**0.03**	**36.45**
HDNet-lite	**0.89M**	**0.02**	**36.91**

Table 5. Average processing time on the CPU.

age harmonization applications.

4.1. Experiment Setting

Datasets. Following the recent works [4, 6, 21], we conduct image harmonization tasks on iHarmony4 benchmark [6]. 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 foreg-round mask, and a composite image (with the fore-ground generated by GAN [9]). We follow the same partition settings of iHarmony4 as DoveNet [6]. Note that we conduct high-resolution (i.e., 1024 × 1024) 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 [25] to implement our models with Nvidia Tesla A100 GPUs.
Evaluation. During the test phase, we use Mean Square Error (MSE), fore-ground 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 6 state-of-the-art methods, including DIH [31], S^2AM [7], DoveNet [6], RainNet [21], Bargainnet [4], Intrinsic [11], D-HT [10], CDTNet [5], Harmonizer [15], DCCF [34] and S^2CRNet [18].

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 ECCV'22 method S^2CRNet-VGG [18] on iHarmony4 dataset, our HDNet brings 12.16 improvement in terms of MSE, and 1.4 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 by 4 times. Compared to the method with equivalent performance, our HDNet-lite has fewer parameters. For example, compared to RainNet and BarginNet, HDNet-lite only uses 2% of the parameters to achieve better performance in the PSNR metric, demonstrating the effectiveness of the proposed network.
Influence of fore-ground ratios. Following [21], we examine the influence of different fore-ground 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 3. It can be observed that our HDNet achieves the best performance

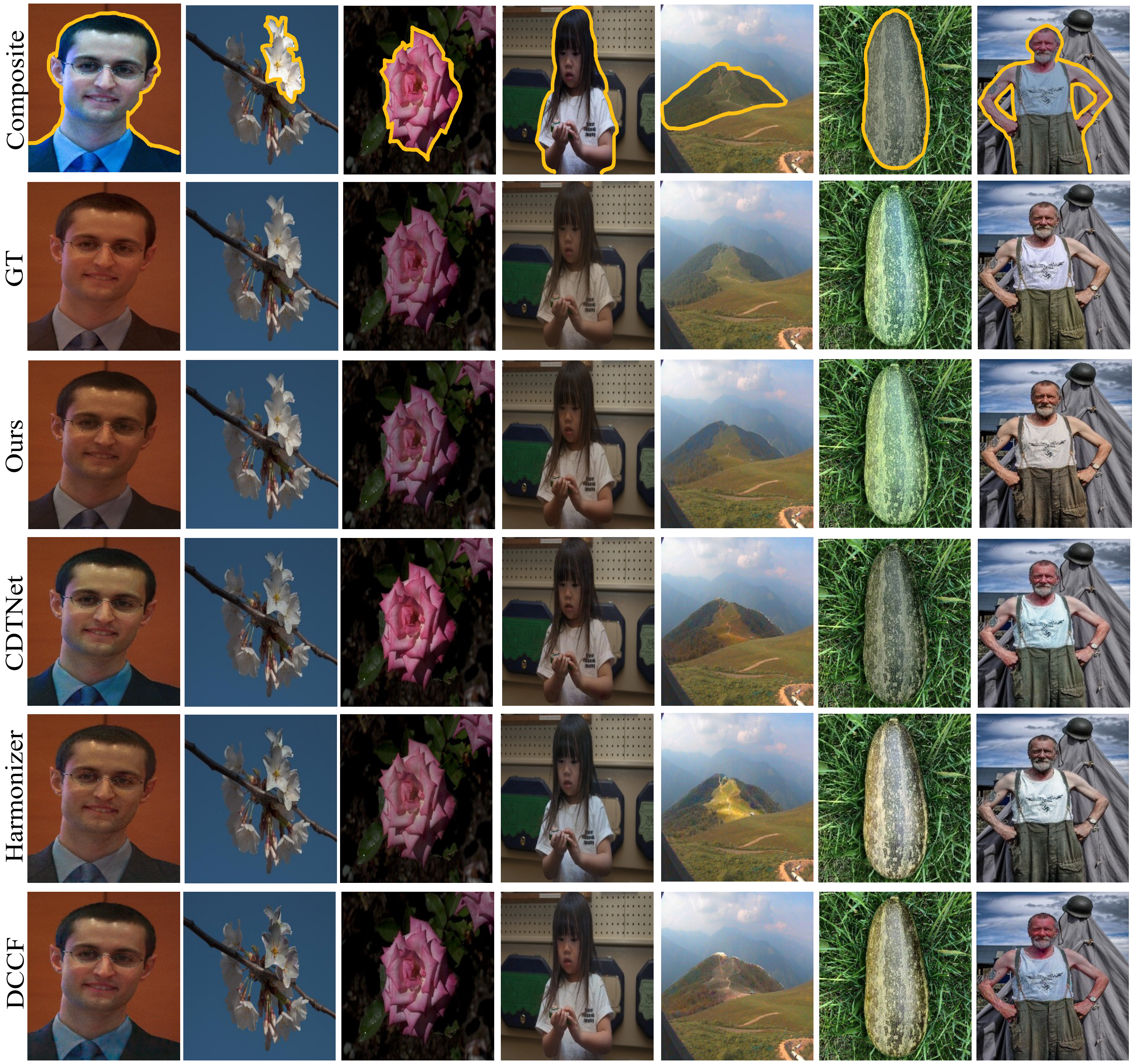


Figure 3. Qualitative comparison on samples from the testing dataset of iHarmony4. The yellow border lines indicate the fore-ground.

among all approaches. HDNet works well at various foreground scales, thanks to its combination of hierarchical dynamics.

High-resolution results. Following [5], we conduct high-resolution image harmonization experiments. As shown in Table 4, we can see that our method outperforms all of them across all metrics. Compared with the most recent method CDTNet [5], our method achieves a huge average performance gain of 5.82 in MSE, 49.05 in fMSE, 0.0031 in SSIM,and 2.49 in PSNR. More high-resolution results can be found in the supplementary.

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 fore-ground objects into the back-ground image, achieving much better visual consistency than other methods. Our HDNet can achieve these photorealistic results because our HDNet adaptively adjusts the feature of fore- and back-ground by hierarchical dynamics learning.

4.3. Ablation Study

Effectiveness of local dynamics. Our local dynamics (LD) module enables features in different channels and different

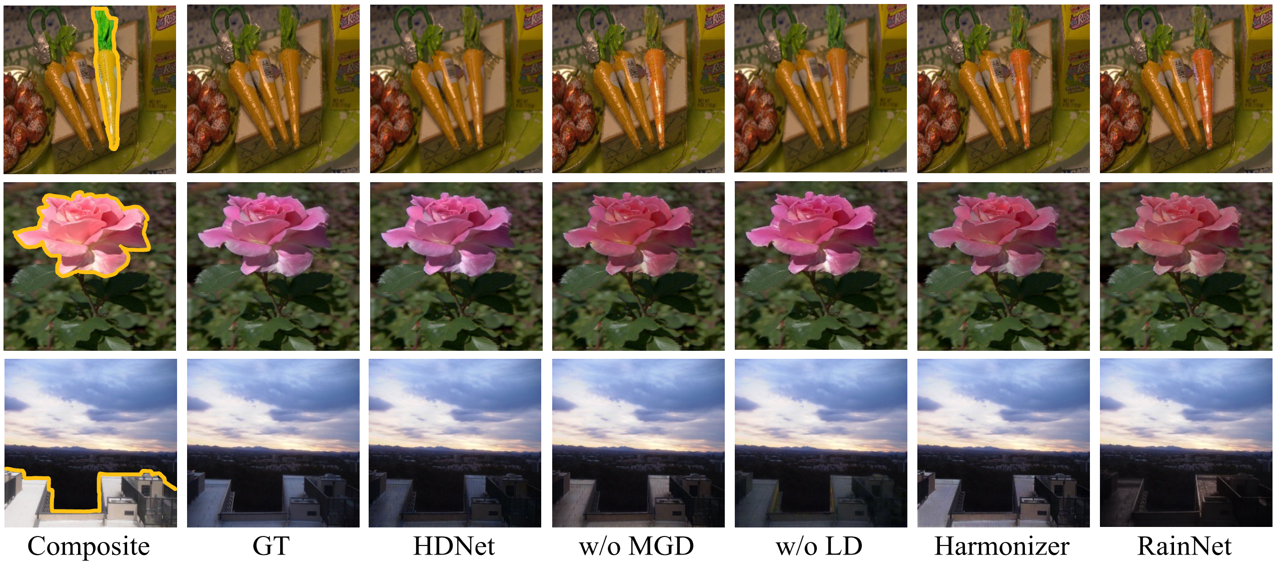


Figure 4. Ablation study on samples from the testing dataset of iHarmony4.

locations varied adaptively by CBAM-like attention. In Table 3, we can see that adding LD to the baseline brings 1.08 dB and 4.65 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 2.77.

Effectiveness of mask-aware global dynamics. Our mask-aware global dynamics (MGD) module integrates the local information to model local coherence. Adding MGD to the model will bring significant improvement, proving that it is not effective enough to perform spatial attention 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 the original mask, the performance will decline, indicating that the learned mask is unreliable.

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.

4.4. Harmonization Performance on CPU

Our HDNet shows good 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 [4, 6, 7, 10, 18] 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 fastest inference speed when inference on CPU and the performance of our model is still better than other methods.

4.5. Limitation

Our method still suffers from some limitations. Our HDNet performs dynamics under the guidance of the 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.

Potential negative impact. Image harmonization is used to create realistic composite images, which may be used to create forged documents and fake news. However, this task does not tamper with the identity of the person or object.

5. Conclusion

This paper proposes a hierarchical dynamic network (HDNet) from local to global that gradually builds local dynamics and mask-aware global dynamics. Local dynamics enable features in different channels and different locations varied adaptively and take advantage of the local consistency of structures to improve the representation ability of geometric transformations; mask-aware global dynamics integrate the local information for modeling the local coherence. Our method achieves state-of-the-art performances on the benchmark dataset iHarmony4 and our lightweight version model HDNet-lite achieves competitive results compared to other methods while only using 2% parameters.

References

[1] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, pages 324–333, 2021. 1

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

[3] Jin Chen, Xijun Wang, Zichao Guo, Xiangyu Zhang, and Jian Sun. Dynamic region-aware convolution. In *CVPR*, pages 8064–8073, 2021. 2, 3, 4

[4] Wenyan Cong, Li Niu, Jianfu Zhang, Jing Liang, and Liqing Zhang. Bargainnet: Background-guided domain translation for image harmonization. In *ICME*, pages 1–6, 2021. 2, 4, 5, 6, 8

[5] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *CVPR*, pages 18470–18479, 2022. 2, 4, 5, 6, 7

[6] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, pages 8391–8400, 2020. 2, 5, 6, 8

[7] Xiaodong Cun and Chi-Man Pun. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Trans. Image Process.*, 29:4759–4771, 2020. 4, 5, 6, 8

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2, 3

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014. 6

[10] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *ICCV*, pages 14850–14859, 2021. 2, 5, 6, 8

[11] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, pages 16367–16376, 2021. 2, 5, 6

[12] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021. 3

[13] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 2, 3, 4

[14] Yongcheng Jing, Xiao Liu, Yukang Ding, Xinchao Wang, Errui Ding, Mingli Song, and Shilei Wen. Dynamic instance normalization for arbitrary style transfer. In *AAAI*, pages 4369–4376, 2020. 3

[15] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *ECCV*, pages 690–706, 2022. 2, 4, 5, 6

[16] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *NeurIPS*, 30, 2017. 3

[17] Yaohui Li, Yuzhe Yang, Huaxiong Li, Haoxing Chen, Liwu Xu, Leida Li, Yaqian Li, and Yandong Guo. Transductive aesthetic preference propagation for personalized image aesthetics assessment. In *ACM MM*, 2022. 3

[18] Jingtang Liang, Xiaodong Cun, and Chi-Man Pun. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In *ECCV*, 2022. 2, 5, 6, 8

[19] Yudong Liang, Bin Wang, Wenqi Ren, Jiaying Liu, Wenjian Wang, and Wangmeng Zuo. Learning hierarchical dynamics with spatial adjacency for image enhancement. In *ACM MM*, pages 2767–2776, 2022. 2

[20] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *NeurIPS*, 34:16331–16345, 2021. 1

[21] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *CVPR*, pages 9361–9370, 2021. 2, 3, 4, 5, 6

[22] J. Matías Di Martino, Gabriele Facciolo, and Enric Meinhardt-Llopis. Poisson image editing. *Image Process. Line*, 6:300–325, 2016. 2

[23] Hyeonseob Nam and Hyo-Eun Kim. Batch-instance normalization for adaptively style-invariant neural networks. In *NeurIPS*, pages 2563–2572, 2018. 3

[24] Dae Young Park and Kwang Hee Lee. Arbitrary style transfer with style-attentional networks. In *CVPR*, pages 5880–5888, 2019. 3

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 6

[26] Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21(5):34–41, 2001. 2

[27] Jing Shi, Ning Xu, Haitian Zheng, Alex Smith, Jiebo Luo, and Chenliang Xu. Spaceedit: Learning a unified editing space for open-domain image color editing. In *CVPR*, pages 19730–19739, 2022. 1

[28] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, pages 1620–1629, 2021. 4

[29] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization. *ACM Trans. Graph.*, 29(4):1–10, 2010. 2

[30] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Not all voxels are equal: Semantic scene completion from the point-voxel perspective. In *AAAI*, volume 36, pages 2352–2360, 2022. 1

[31] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 2799–2807, 2017. 2, 5, 6

[32] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, pages 11531–11539, 2020. 4

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 2, 3, 4

[34] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. DCCF: deep comprehensible color filter learning framework for high-resolution image harmonization. In *ECCV*, 2022. 5, 6

[35] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly E. Rushmeier. Understanding and improving the realism of image composites. *ACM Trans. Graph.*, 31(4):84:1–84:10, 2012. 2

[36] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets V2: more deformable, better results. In *CVPR*, pages 9308–9316, 2019. 3