2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.00693

Learning Global-aware Kernel for Image Harmonization

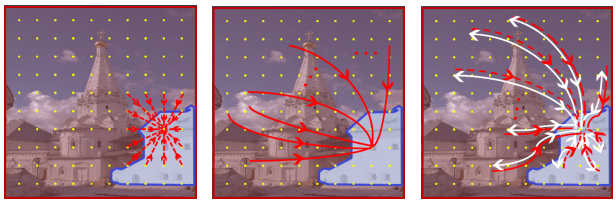
Xintian Shen[1]* Jiangning Zhang[2]* Jun Chen[1] Shipeng Bai[1]
Yue Han[1] Yabiao Wang[2] Chengjie Wang[2,3] Yong Liu[1]†
[1] APRIL Lab, Zhejiang University [2]Youtu Lab, Tencent [3]Shanghai Jiao Tong University
[22132133, 186368, junc, shipengbai, 22132041]@zju.edu.cn,
[caseywang, jasoncjwang]@tencent.com, yongliu@iipc.zju.edu.cn

Abstract

Image harmonization aims to solve the visual inconsistency problem in composited images by adaptively adjusting the foreground pixels with the background as references. Existing methods employ local color transformation or region matching between foreground and background, which neglects powerful proximity prior and independently distinguishes fore-/back-ground as a whole part for harmonization. As a result, they still show a limited performance across varied foreground objects and scenes. To address this issue, we propose a novel Global-aware Kernel Network (GKNet) to harmonize local regions with comprehensive consideration of long-distance background references. Specifically, GKNet includes two parts, i.e.*, harmony kernel prediction and harmony kernel modulation branches. The former includes a Long-distance Reference Extractor (LRE) to obtain long-distance context and Kernel Prediction Blocks (KPB) to predict multi-level harmony kernels by fusing global information with local features. To achieve this goal, a novel Selective Correlation Fusion (SCF) module is proposed to better select relevant long-distance background references for local harmonization. The latter employs the predicted kernels to harmonize foreground regions with local and global awareness. Abundant experiments demonstrate the superiority of our method for image harmonization over state-of-the-art methods,* e.g.*, achieving 39.53dB PSNR that surpasses the best counterpart by +0.78dB ↑; decreasing fMSE/MSE by 11.5%↓/6.7%↓ compared with the SoTA method. Code will be available at here.*

1. Introduction

Image composition aims to synthesize foreground objects from one image into another, which is a common task in image editing. However, human eyes could clearly distinguish synthetic images due to the visual inconsistency between foreground and background in composited images. In attempting to solve the photo-unrealistic problem, image harmonization is proposed to adjust the foreground objects based on the illumination and color tone in background environment, which plays an important role in image editing.



(a) Local-translation (b) Region-matching (c) Ours

Figure 1. Comparison of background reference methods in harmonization. Blue/Red region represent foreground/background, respectively, and white/red arrows refer to interaction/injection, respectively. (a) Local-translation methods reference nearby pixels. (b) Region-matching methods transfer reference with a unified view of fore-/back-ground region. (c) Our method interacts long-distance reference and injects it with short-distance consideration.

Traditional image harmonization approaches are mainly based on low-level feature matching, which is only effective for specific scenes. Recently, numerous learning-based methods have achieved remarkable progress by addressing image harmonization as a generation task. Existing learning-based methods could be categorized from two angles, *i.e.*, *local-translation* and *region-matching*. 1) The former employs a convolutional encoder-decoder to learn a foreground pixel-to-pixel translation [41, 8]. But a shallow CNN only captures limited surrounding background. As shown in Figure 1a, these approaches harmonize the current pixel with local references, which is insufficient for harmonization as inner foreground pixels could not attach background reference. Besides, related long-distance references are effective in some cases. 2) The latter region matching methods [29, 7] distinguish foreground and background regions as two styles or domains. As shown in Figure 1b, they tackle harmonization as a matching problem with a

*Equal contribution.
†Corresponding author.

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.00693

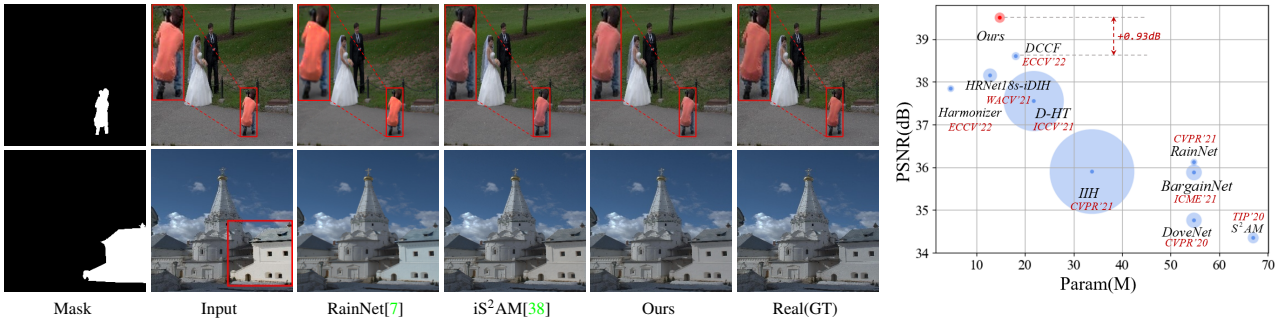


Figure 2. **Left:** Two challenging samples in image harmonization. Mask in column one and Red boxes represents the foreground. **Right:** Performance comparison with SOTA methods in terms of PSNR and model size. The circle size represents the floating-point number.

unified view of these two regions by statistics components or discriminators. Though these approaches harmonize images with a broader reference range, they totally neglect the spatial variations in two regions. Hang *et al*. [17] begin to notice this problem and add attention-based references in region matching method [29]. But they still separate two regions independently and harmonize foreground by unified matching without considering foreground spatial variations.

To further illustrate existing problems, we provide two common harmonization cases in Figure 2. In the first case, a small foreground object appears in the background with obvious color changes. The region-matching method RainNet [7] provides a poor color correction result while the local method iS^{2}AM [38] could tackle this case well, which indicates that the unified view of background will blend the overall complex color conditions. In the second case, related long-distance references exist in the background, while the local method could only attach insufficient adjacent information. Region-matching method RainNet could obtain whole background blue tone by matching, but it still excessively harmonizes the house due to the unified view. These two cases indicate that local reference is insufficient, but region-matching methods could not model long-distance reference well and will cause unbalanced harmonization problems by rough matching.

To solve this problem, we rethink essential proximity priors in image harmonization, *i.e*., when we paste an object into background, the color or light is related to location and will be influenced by its neighboring first. Moreover, the effective long-distance information in background changes with pasted locations, which requires us to learn adaptive references for each part. Inspired by this observation, we propose a novel *Global-aware Kernel Network* (GKNet) to integrate local harmony modulation and long-distance background references, including harmony kernel prediction and harmony kernel modulation. For harmony kernel prediction, we propose a novel global-aware kernel prediction method including *Long-distance Reference Extractor* (LRE) to obtain long-distance references and *Kernel Prediction Blocks* (KPB) to predict multi-level adaptive kernels with selected long-distance references by *Selective Correlation Fusion* (SCF). For kernel modulation, we propose to model local harmony operation by predicted global-aware kernels and multi-level features. Focusing on features in kernel region, kernel modulation is significant in alleviating unbalanced region-matching errors in complex scenes.

To summarize, we make following contributions:

- With the observation of proximity prior and long-distance references in image harmonization task, we design GKNet to model global-local interaction by learning global-aware harmony kernel, including harmony kernel prediction and harmony kernel modulation.
- For harmony kernel prediction, we propose a kernel prediction branch combined with LRE to model global information and multiple KPB to learn adaptive harmony kernels. For better global references, we design SCF to select relevant long-distance references for local harmonization. For harmony kernel modulation, we propose the method to harmonize local regions in multi-level decoder layers with predicted kernels.
- Extensive experiments demonstrate the superior performance of our methods in both quantitative and qualitative results, noting that our method achieves state-of-the-art results on iHarmony4 datasets.

2. Related work

2.1. Image Harmonization

Traditional image harmonization works have focused on finding a better method for low-level appearances matching between foreground and background regions in images, which includes color statistics [36, 35, 45], gradient information [20, 34, 40], and multi-scale statistical features [39, 25]. However, traditional methods could only be effective in specific scenes. With the advanced generative ability of deep learning, Tsai *et al*. [41] firstly propose a learning-based encoder-decoder network assisted by

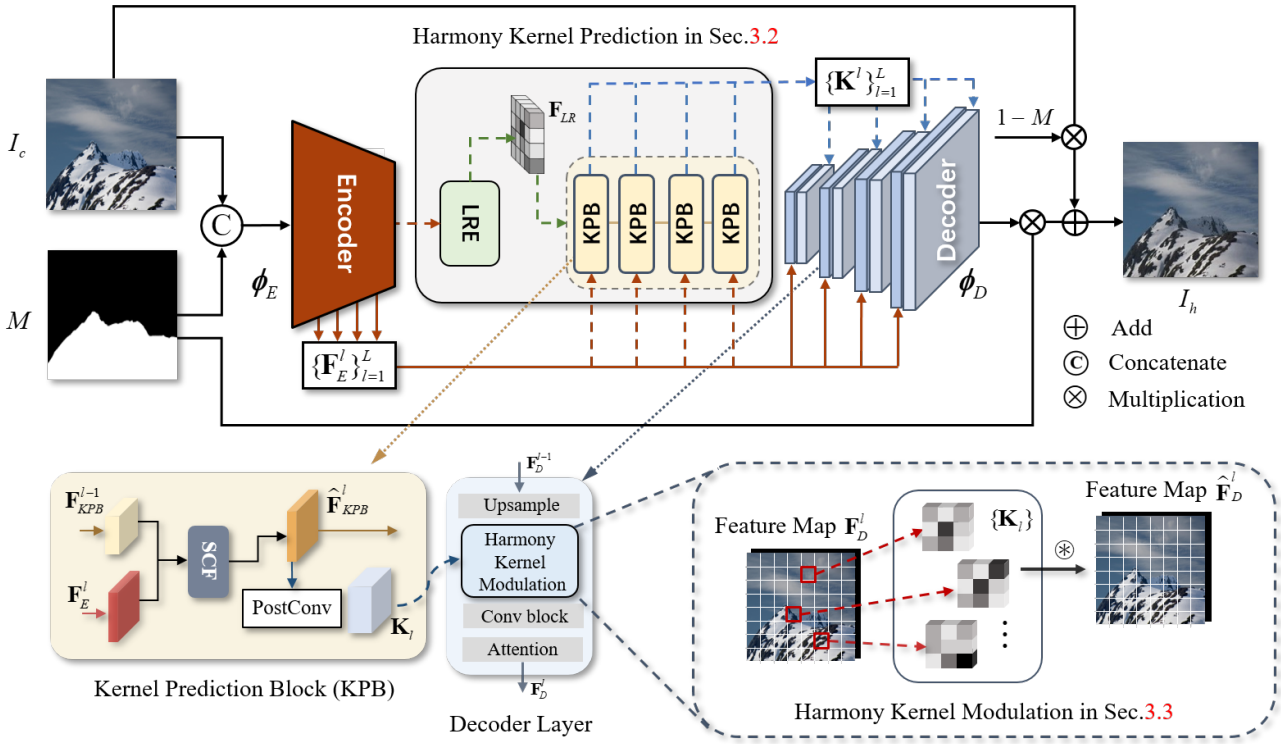


Figure 3. **The overview of our proposed GKNet**, which consists of harmony kernel prediction branch and harmony kernel modulation branch. As shown in gray box, the harmony kernel prediction branch is combined with a *Long-term Reference Extractor* (LRE) and multi-level *Kernel Prediction Blocks* (KPB). As shown in yellow box, we propose *Selective Correlation Fusion* (SCF) module in KPB for better long-distance references. Given a composited image I_c with corresponding foreground mask M, we extract deep features $\mathbf{F}_E^l$ from encoder ϕ_E. Then, harmony kernel prediction branch utilizes the deepest feature map and $\{\mathbf{F}_E^l\}$ to predict multi-level dynamic harmony kernels $\{\mathbf{K}^l\}$ increasingly. The predicted global-aware kernels are employed for harmony kernel modulation in decoder ϕ_D.

a semantic branch. In observation that semantic information is effective in image harmonization, Soffiuk *et al*. [38] also add additional pre-train semantic model to baseline DIH [41] and S^2AM [8]. Inspired by domain transfer, Cong *et al*. [7] adopt a verification discriminator to distinguish foreground and background domains. Similarly, Ling *et al*. [29] also treat the composited image as two independent parts and apply style transfer idea to match mean-variance statistics. To focus on harmonize foreground region, some methods add attention mechanisms. Cun *et al*. [8] add a spatial-separated attention module. Guo *et al*. [15] for the first time introduce Transformer architecture to image harmonization. Hang *et al*. [17] add background attention calculation to the style transfer block [29], and they also incorporated the idea of contrast learning. Besides, Guo*et al*. [16] decompose image into reflectance and illumination by autoencoder for separate harmonization based on Retinex theory. Some high-resolution methods [24, 44] frame image harmonization as an image-level problem to learn white-box arguments. However, the *above methods neglect spatial proximity prior and could not model long-distance references well*. Instead in this paper, we design a better local modulation method combined with selected long-distance references to alleviate this problem.

2.2. Dynamic Filtering in Image Editing

The input-dependent dynamic filtering first proposed by Jia [21] *et al*. aims to learn position-specific filters on pixel inputs and apply the generated kernels to another input, which has been widely used in numerous vision tasks [33, 23, 46, 9]. This method also shows effectiveness in image editing tasks, such as denoising [1, 32, 42], shadow removing [13], deraining [14], image inpainting [26], and blur synthesis [2]. However, most above methods apply dynamic filtering at image-level filter prediction and utilization. We propose to learn a multi-level global-aware kernel with long-term context references for harmonization.

2.3. Feature Fusion

Feature fusion is to combine features from different layers or branches, which is an omnipresent part of modern neural networks and has been studied extensively. Most previous works [18, 37, 28] for feature fusion focus on the pathways structure design, applying two linear classic

methods of summation or concatenation. Recently, benefit from the successful use of Transformer in computer vision [4, 11, 31, 50, 49, 48, 3, 4, 12, 43, 22, 30, 15], some works [19, 27, 52, 10, 47, 51, 5] apply attention mechanism to present nonlinear approaches for feature fusion. As global information is significant in image harmonization, we design a dynamic weighted fusion method to effectively fuse long-distance reference into kernel prediction.

3. Method

Given a composited image I_c with its corresponding binary mask M indicating the region to be harmonized, our goal is to learn a network G that outputs harmonized image I_h, which could be formulated as $I_h = G(I_c, M)$. To make this composited image I_c look natural, we train our model G to adjust the foreground region I_f in a supervised manner with paired real image I. In this paper, we also define the background region as I_b, and then the composition process could be formulated as $I_c = I_f \cdot M + (1 - M) \cdot I_b$, where $\cdot$ denotes element-wise multiplication.

3.1. Overview of Our Network

As shown in Figure 3, we design a novel network architecture for image harmonization tasks to allow our network to pay attention to short-distance and long-distance information simultaneously. Following the standard designs in image harmonization works [8, 38, 29] we use simple U-Net [37] with attention blocks [8] as the basic structure. We also take composited RGB image $I_c \in \mathbb{R}^{3\times H\times W}$ concatenated with foreground region mask $\mathrm{M} \in \mathbb{R}^{1\times H\times W}$ as input.

Motivated by proximity prior in image harmonization, we propose *Global-aware Kernel Network* (GKNet) to learn global-aware harmony kernel for image harmonization, which consists of two branches, harmony kernel prediction and harmony kernel modulation. Firstly, as long-distance reference is crucial for harmonization task, we design global-aware kernel prediction branch to predict harmony kernel with context modeling, which contains a transformer-based *Long-term Reference Extractor* (LRE) to extract global reference and *Kernel Prediction Block* (KPB) to predict harmony kernels. In order to incorporate relevant long-term references for local harmonization, a novel *Selective Correlation Fusion* (SCF) is proposed to select more effective references in backgrounds. Secondly, we design a multi-level harmony kernel modulation in decoder layers to employ the predicted global-aware kernels. The mechanism between global-aware harmony kernel prediction and harmony kernel modulation finally achieves local-global interaction for image harmonization.

3.2. Harmony Kernel Prediction

Inspired by recent works in image editing [1, 14, 32], we propose to apply dynamic kernels to force the current pixel harmonized with surrounding regions adaptively. This approach effectively makes up for the lack of consideration of proximity prior in previous harmonization works. However, basic dynamic kernels for image editing tasks such as denoising and deraining are applied with a fixed size at image-level. To predict more proper adaptive kernels for image harmonization, we analyze the following: **1)** Global modeling is necessary for image harmonization as long-distance references may appear in the background. Hence, we design a novel global-aware harmony kernel prediction branch with LRE to extract global information and KPB to predict global-aware kernels with fusion module SCF. **2)** Fixed-size kernels applied at image-level could not handle the scale problem well. *e.g.* The pixels inside the large foreground mask can hardly obtain any real background information, while predicting large kernels will bring high computation costs and breaks the intention of proximity prior. Besides, image-level dynamic kernels pay more attention to detailed structure, while in image harmonization we also need to adapt multiple scene variations to harmonize foregrounds at semantic level. In order to adapt to multi-scene and multi-scale problems, we propose to predict kernels in multi-level structures.

Long-distance Reference Extractor. In order to obtain long-term context, we employ l-transformer layers [11] as our global information extractor. We feed the deepest feature map $\mathbf{F}_E^1$ from CNN encoder into transformer layers. With the down-sampling feature map in low-resolution of $(\frac{w}{r}, \frac{h}{r})$, we treat each pixel as a token to generate embeddings. With the multi-head attention mechanism, we obtain global interactive feature $\mathbf{F}_{global} \in \mathbb{R}^{C\times HW}$ after l-transformer layers. After reshaping and post-convolution layer, we obtain the long-term reference feature $\mathbf{F}_{LR}$.

Kernel Prediction Block. To adapt diverse foreground scales and background scenarios, we apply our local operation kernel modulation in multiple decoder levels. Nevertheless, deep-level features contain more semantic information, and shallow features contain more details, we need to predict corresponding adaptive kernels for different level harmony kernel modulation. Thus, our designed global-aware harmony kernel prediction branch is in a multi-level structure to increasingly predict a series of kernels.

In Figure 3, we show our proposed KPB structure from predicting kernels to harmony kernel modulation operation. The operation in lth KPBlock can be formulated as

$$\mathbf{K}^l, \mathbf{F}_{KPB}^l = KPBlock(\mathbf{F}_E^l, \mathbf{F}_{KPB}^{l-1}), \tag{1}$$

where KPBlock$(\cdot)$ is the KPBlock to predict $\mathbf{K}^l$. For each KPB, we take $\mathbf{F}_{KPB}^{l-1} \in \mathbb{R}^{C_l\times H_l\times W_l}$ transferred from $(l-1)$th KPB (For the deepest KPB, we input $\mathbf{F}_{LR}$) and the lth encoder layer feature $\mathbf{F}_E^l$ as input (We denote $\mathbf{F}_E^1$ as the deepest

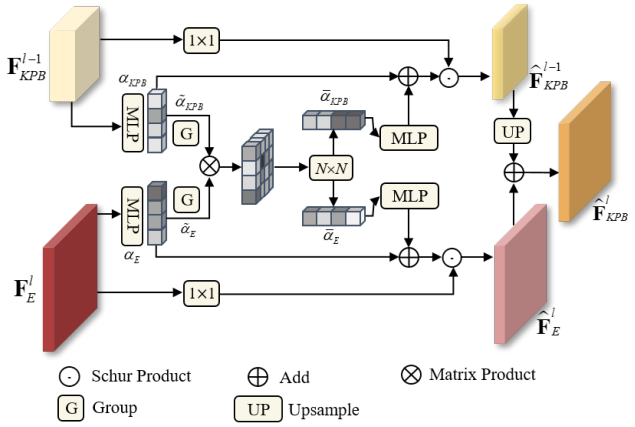


Figure 4. **Schematic diagram of SCF.** The module takes the (l-1)th layer feature $\mathbf{F}^{l-1}_{KPB}$ and encoder feature $\mathbf{F}^{l}_{E}$ as input and outputs the correlation-aware fusion feature $\hat{\mathbf{F}}^{l}_{KPB}$.

layer in encoder), which outputs feature $\mathbf{F}^{l}_{KPB}$ for next KPBlock and the global-aware kernels $\mathbf{K}^{l} = \text{Conv}(\mathbf{F}^{l}_{KPB})$ after post-convolutions.

Selective Correlation Fusion. Long-distance reference $\mathbf{F}_{LR}$ obtained from LRE is then injected into the deepest KPBlock to model global information for local harmonization. The standard way of feature fusion, like concatenation or addition, equally treats low-level and high-level features. To efficiently model long-term information for local harmonization, we propose SCF to select relevant global information by interacting encoder features $\mathbf{F}_E$ and long-distance references based on channel-wise attention mechanism.

Specifically, as shown in Figure 4, we take the $(l-1)$th layer feature $\mathbf{F}^{l-1}_{KPB} \in \mathbb{R}^{C_{l-1}\times H_{l-1}\times W_{l-1}}$ and the lth encoder feature $\mathbf{F}^{l}_{E} \in \mathbb{R}^{C_l\times H_l\times W_l}$ as input and extract attention vector $\alpha_{KPB}, \alpha_E \in \mathbb{R}^{C_l}$ by 3×3 convolutions and MLP. Subsequently, the attention vector is divided into n groups with length m as $\tilde{\alpha}_{KPB}, \tilde{\alpha}_E \in \mathbb{R}^{n\times m}$. Thus, the channel-wise attention relation $\mathbf{A} \in \mathbb{R}^{n\times n}$can be calculated by matrix product

$$\mathbf{A} = \tilde{\alpha}_{KPB} \odot \tilde{\alpha}^{T}_{E}. \tag{2}$$

After that, we calculate the selective factor $\overline{\alpha}_{KPB}, \overline{\alpha}_E \in \mathbb{R}^{C_l}$ through $N\times N$ convolutions with splitting. Then, we obtain the selective attention weights $S^l \in \{S^l_E, S^{l-1}_{KPB}\}$ for each features by $\alpha \in \{\alpha_{KPB}, \alpha_E\}$ and $\overline{\alpha} \in \{\overline{\alpha}_{KPB}, \overline{\alpha}_E\}$, which can be formulated as

$$S^l = \sigma(\alpha + b \cdot FC(\overline{\alpha})), \tag{3}$$

where b is a learnable parameter, σ is sigmoid function. Based on the attention weight vector, then the shallow and deep information are interacted as follows:

$$\hat{\mathbf{F}}^{l}_{KPB} = S^l_E \cdot Conv(\mathbf{F}^{l}_{E}) + \text{Upsample}(S^{l-1}_{KPB} \cdot Conv(\mathbf{F}^{l-1}_{KPB})), \tag{4}$$

where $\cdot$ denotes to element-wise multiplication.

3.3. Harmony Kernel Modulation

The global-aware adaptive kernel obtained from the harmony kernel prediction branch is then utilized in harmony kernel modulation. In this section, we illustrate the regional harmony operation method kernel modulation in decoder layers, which converts the previous overall treatment of background and foreground into a local reference. As mentioned in Section 3.2, we apply multi-level kernel modulation in decoder layer to adapt scales and scenario variation problem. Figure 3 shows our proposed kernel modulation in lth decoder layers, which could be formulated as

$$\hat{\mathbf{F}}^{l}_{D} = \mathbf{F}^{l}_{D} \circledast \mathbf{K}^{l}, \tag{5}$$

where $\circledast$ denotes the harmony kernel modulation, $\mathbf{F}^{l}_{D} \in \mathbb{R}^{C\times H\times W}$ is the deep feature extracted from the lth layer in decoder, and the $\hat{\mathbf{F}}^{l}_{\mathbf{D}} \in \mathbb{R}^{C\times H\times W}$ is its corresponding harmony kernel modulation result feature in the lth layer. The tensor $\mathbf{K}^{l} \in \mathbb{R}^{C\times N^2\times H\times W}$ represents the kernels with size of N for harmony kernel modulation in the lth feature layer, which we obtain from KPB. For the kernel modulation in each pixel, we can expand the above equation as

$$\hat{\mathbf{F}}^{l}_{D}[\mathbf{p}] = \sum_{\mathbf{q}\in\mathcal{N}_p} \mathbf{K}^{l}_{p}[\mathbf{p}-\mathbf{q}]\mathbf{F}^{l}_{D}[\mathbf{q}], \tag{6}$$

where $\mathbf{p}$ and $\mathbf{q}$ are the coordinates of pixels in the image, $\mathbf{K}^{l}_{p}$ is the kernel for filtering the element $\mathbf{p}$ of $\mathbf{F}^{l}_{D}$ via its surrounding elements, *i.e.*,$\mathcal{N}_p$. As we illustrate in Eq. 5, $\mathbf{K}^{l}$ contains all element-wise kernels,*i.e.*, $\mathbf{K}^{l}_{p} \in \mathbb{R}^{C\times N\times N}$ for filtering operations. After the kernel modulation in decoder layers, we finally obtain the harmonization result. In this paper, we define the U-Net decoder layer as $\phi_D(\cdot)$, then we can formulate our harmonization process with kernel modulation as $\hat{I}_h = \phi^{L}_{D}(\cdots\phi^{1}_{D}(\mathbf{F}^{1}_{D} \circledast \mathbf{K}^{1}))\cdot M + (1-M)\cdot I_c$.

3.4. Objective function

In the training phase, we only employ foreground-normalized MSE loss as our objective function. Compared with normal MSE loss, it reduces the impact of copying background area:

$$\mathcal{L}_{rec} = \frac{\sum_{h,w} \left\|\hat{I} - I\right\|^2_2}{\max\left\{A_{min}, \sum_{h,w} M_{h,w}\right\}}, \tag{7}$$

where A_{min} is a hyperparameter to keep the loss function stable as there might be some too small foreground objects. In this paper, we set $A_{min} = 100$ as suggested in [38].

4. Experiments

4.1. Implementation Details

We conduct the image harmonization experiment at resolution 256×256 on the benchmark dataset iHarmony4 [7].

Table 1. Quantitative comparisons across four sub-datasets of iHarmony4 [7]. ↑ indicates the higher the better, and ↓ indicates the lower the better. We compute fMSE for better reflection on harmonization tasks. Best results are in bold and the suboptimal results are in underline.

Method	HCOCO			HAdobe5k			HFlickr			Hday2night			ALL		
	PSNR↑	MSE↓	fMSE↓	PSNR↑	MSE↓	fMSE↓	PSNR↑	MSE↓	fMSE↓	PSNR↑	MSE↓	fMSE↓	PSNR↑	MSE↓	fMSE↓
Composite	33.94	69.37	996.59	28.16	345.54	2051.61	28.32	264.35	1574.37	34.01	109.65	1409.98	31.63	172.47	1376.42
DIH [41]	34.69	51.85	798.99	32.28	92.65	593.03	29.55	163.38	1099.13	34.62	82.34	1129.40	33.41	76.77	773.18
S^2AM [8]	35.47	41.07	542.06	33.77	63.40	404.62	30.03	143.45	785.65	35.69	50.87	835.06	34.35	59.67	594.67
DoveNet [7]	35.83	36.72	551.01	34.34	52.32	380.39	30.21	133.14	827.03	35.27	51.95	1075.71	34.76	52.33	532.62
IIH [16]	37.16	24.92	416.38	35.20	43.02	284.21	31.34	105.13	716.60	35.96	55.53	797.04	35.90	38.71	400.29
RAINNet [29]	37.08	29.52	501.17	36.22	43.35	317.35	31.64	110.59	688.40	34.83	57.40	916.48	36.12	40.29	469.60
iDIH-HRNet [38]	39.16	16.48	266.19	38.08	21.88	173.96	33.13	69.67	443.65	37.72	40.59	590.97	38.19	24.44	264.96
D-HT [15]	38.76	16.89	299.30	36.88	38.53	265.11	33.13	74.51	515.45	37.10	53.01	704.42	37.55	30.30	320.78
Harmonizer [24]	38.77	17.34	298.42	37.64	21.89	170.05	33.63	64.81	434.06	37.56	**33.14**	**542.07**	37.84	24.26	280.51
DCCF [44]	39.72	14.55	267.79	38.24	20.20	171.01	33.72	66.20	440.84	38.18	51.40	629.67	38.60	22.64	265.41
SCS-Co [17]	39.88	13.58	245.54	38.29	21.01	165.48	34.22	**55.83**	393.72	37.83	41.75	606.80	38.75	21.33	248.86
Ours	**40.32**	**12.95**	**222.31**	**39.97**	**17.84**	**138.22**	**34.45**	57.58	**372.90**	**38.47**	42.76	546.06	**39.53**	**19.90**	**220.44**

Table 2. Quantitative comparisons on different ratios of foreground based on iHarmony4 by MSE and fMSE metrics. The best results are in bold and the suboptimal results are in underline.

Method	Venue	0%~5%		5%~15%		15%~100%	
		MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
DIH	CVPR'17	18.92	799.17	64.23	725.86	228.86	768.89
S^2AM	TIP'20	15.09	623.11	48.33	540.54	117.62	592.83
DoveNet	CVPR'20	14.03	591.88	44.90	504.42	152.07	505.82
RainNet	CVPR'21	11.66	550.38	32.05	378.69	117.41	389.80
iS^2AM	WACV'21	6.73	294.76	18.03	204.69	63.02	207.82
GKNet	Ours	**5.36**	**244.06**	**17.46**	**200.34**	**57.31**	**188.75**

The initial learning rate is set to 10^{-4}, and the models are trained for 120 epochs with a batch size of 16 on four 2080Ti GPUs. For optimizer, we adopt an Adam optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-8}$. It takes about two days for training. Our proposed model is implemented by PyTorch, and more detailed network architectures could be found in the supplementary file.

4.2. Datasets and Metrics

Datasets. To evaluate our proposed method for image harmonization, we conduct our experiments on the benchmark dataset iHarmony4 [7], which consists of 4 sub-datasets: HCOCO, HAdobe5K, HFlickr, and Hday2night, including 73147 pairs of synthesized composite images with their corresponding foreground mask and ground truth image.

Evaluation Metrics. Following the standard setups in image harmonization, we use the peak signal-to-noise ratio (PSNR) and Mean Squared Error (MSE) as evaluation metrics. Furthermore, it is more accurate to only calculate the difference in the foreground region with the metric foreground MSE (fMSE) introduced by [38]. The metric MSE is calculated for the whole image, while there is no changes for pixels in the background region in harmonization task. Without considering the foreground ratio, the average MSE and PSNR results of the dataset will be more responsive to the performance of large-scale targets. In this paper, we argue that fMSE is more suitable for harmonization task.

4.3. Comparison with SOTAs

Quantitative comparison. As results shown in Table 1, we compare our method with other state-of-the-art image harmonization methods on iHarmony4 [7]. Following [7, 38], we also evaluate the model performance on different ratios of foreground by splitting the test images into three groups,*i.e.*, 0% ~5%, 5%~15%, and 15%~100%. We provide these results in Table 2. Observing the quantitative experiment results above, we can summarize the following conclusions: **1)** Our method achieves SOTA results of all evaluation metrics on average iharmony4 datasets. More specifically, our method achieves 0.78dB↑ improvement in PSNR, 1.43↓ in MSE, and 28.42↓ in fMSE compared with suboptimal methods. **2)** Our method obtains the best fMSE scores on all sub-datasets, meaning that the foreground regions generated by our method are more natural and closer to real images. **3)** As shown in Table 2, our model performs well on each foreground ratio, especially on 0%~5%, which indicates that our method has a strong ability to handle variable foreground scales. It also proves that our proposed global-aware adaptive kernel could process global-to-local impact with excellence.

Qualitative comparison. We further provide qualitative results on iHarmony4 datasets in Figure 5. It could be observed that our GKNet generates more visually coherent images of the foreground and background than composited images, which are also closer to ground truth. **1)** The first two rows of examples show that our method can effectively handle the small object harmonization problem in complex scenes with proximity prior consideration. **2)** The last two row examples show that our regional modulation with multi-level structure performs well in large foreground cases, which handles a wide range of enormous contrast. For more detailed descriptions, please refer to the caption in Figure 5. In contrast, our method presents more photorealistic results. More visual comparison results on iHarmony4 datasets could be seen in supplementary materials.

Comparisions on Real Datasets. Experiments results on

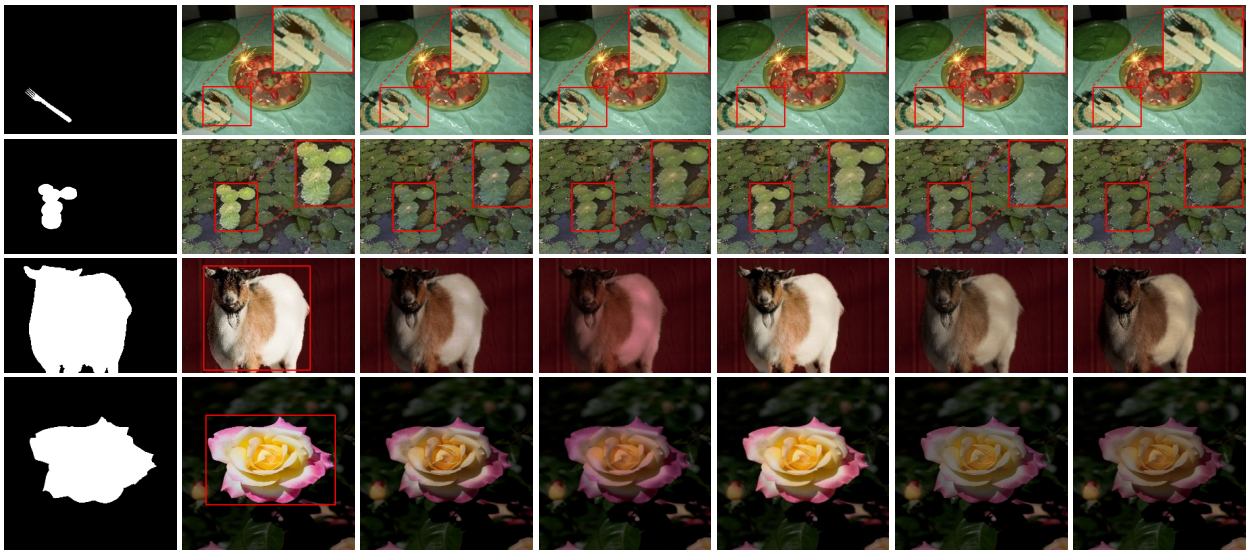


Figure 5. **Qualitative comparisons with SOTA methods on iHarmony4 [7].** Mask in column one and Red box in Input represents the foreground. **Case 1:** The background color shows spatial variation, while only our method captures practical color reference by predicted kernel and local harmony modulation. **Case 2:** The harmony result obtained by our method is better in the detailed structure like duckweed center due to target harmony kernels for each foreground part. **Case 3 & 4:** For large foregrounds, our approach could also achieve better results, which preserves more original details and is closer to ground truth.

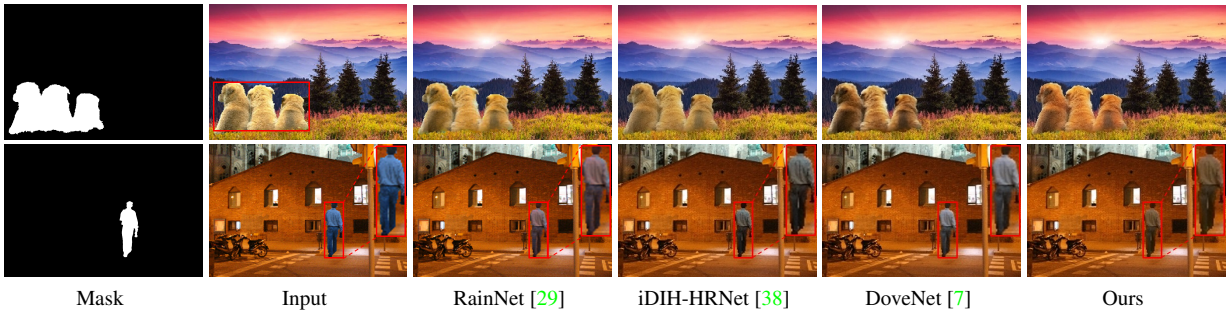


Figure 6. **Qualitative comparisons with SOTA methods on real composite images** [41]. Mask in column one and Red box in Input represents the foregrounds. **Case 1:** Comparing the results on the dogs, our method achieves more natural results over the sunset background. **Case 2:** Our result dims the foreground person based on background dim light, which achieves a more natural effect.

real composition datasets [41] can be seen in Figure 6. For real composition cases, evaluation metrics are impossible to calculate as there are no ground truth images. Hence, we only show qualitative results and human study results here. More visual comparison results on real datasets could be seen in supplementary materials.

Table 3. B-T scores comparison on real composite images.

Method	Composite	DoveNet[7]	RainNet[29]	iDIH-HRNet[38]	Ours
B-T Score ↑	0.416	0.686	0.972	1.532	**1.944**

Following [7, 6, 16, 17], we conduct our human study by inviting 50 volunteers to compare 24750 pairwise results. The pairwise results are obtained from 99 real composited images, with 25 results for each pair of different methods on average. We also use the Bradley-Terry model (B-T model) to calculate the global ranking score. Our method achieves the best results as shown in Table 3.

4.4. Ablation Studies

Effectiveness of network components. We further conduct quantitative experiments to verify the effectiveness of components in GKNet. Note that modules in GKNet have dependencies, we can only show gradually added ablation studies. As the results in Table 4, our full model obtains the highest performance on all metrics when KPB, LRE, and SCF work together. Table 4 also illustrates the effectiveness of each component. Moreover, to further illustrate our global-local interaction method for harmonization, we show

Table 4. Quantitative ablation study of our approach with different components on iHarmony4[7]

Baseline	KPB	LRE	SCF	MSE ↓	PSNR ↑	fMSE ↓
✓	✗	✗	✗	27.27	37.83	280.56
✓	✓	✗	✗	21.42	38.68	235.72
✓	✓	✓	✗	20.50	39.30	229.63
✓	✓	✓	✓	19.90	39.53	220.44

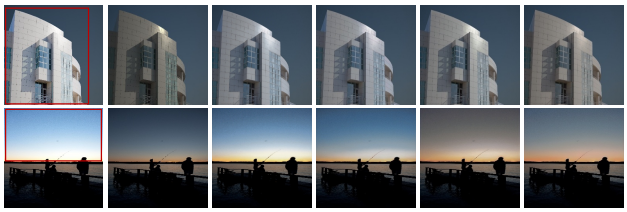


Figure 7. Qualitative ablation study of our approach. Red boxes in input image mark foreground.

qualitative results in Figure 7. Compared with the baseline, the model with only local KPB outputs stronger local modeling results but loses global view. After introducing LRE, global awareness of the background is enhanced.But it is evident that color is overcorrected in the second row. After adding SCF, we effectively correct the deviation by selecting relevant references in background, and the final result is closer to real image.

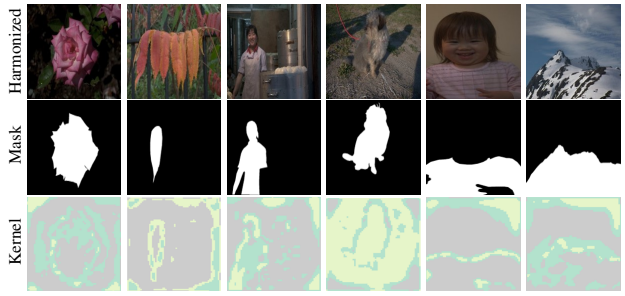
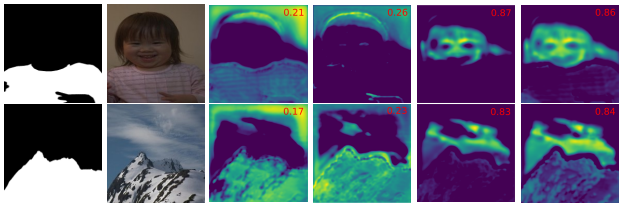


Figure 8. The cluster visual results of adaptive harmony kernels. *e.g.*, in flower case, structure information like petal margin is displayed, which proves that kernels are predicted to adapt local harmonization specifically.

Interpretability of Global-aware Harmony Kernel. To further illustrate the effectiveness of the adaptive harmony kernels, we cluster the per-pixel adaptive harmony kernels predicted from KPB by K-means. As shown in Figure 8, the clusters show strong spatial structure, which indicates that *our predicted dynamic kernel can make the structural adjustment to harmonize the foreground*. Moreover, in some cases, the change of kernel classes is related to the object mask. This exhibits that our harmony kernels are predicted dynamically to deal with visual inconsistency in different spatial locations (*e.g.* fore-/-background or edges).



(a) Mask (b) Comp (c) Feature map with channel attention weight

Figure 9. Feature maps in SCF. Red number in upper right corner of feature map represents channel-wise attention weights.

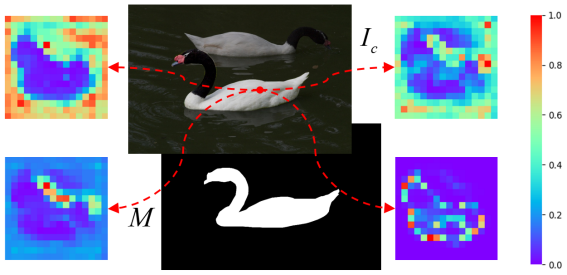


Figure 10. Cross-attention from example point (10,10). We show four attention maps for different heads, which proves LRE can match model related long-distance references.

Interpretability of LRE and SCF Module. We visualize the attention mechanism in LRE and SCF for interpreting global-local interaction. In Figure 9, we visualize feature maps in decoder layers to illustrate our channel-wise attention module SCF. Visual feature maps with attention weights show that harmony kernels are predicted with more attention on related background area and less attention on irrelevant background area or foreground. In Figure 10, we visualize the attention maps of LRE, focusing on an example point in foreground. As the long-distance information in predicted harmony kernels is brought by LRE, the visualized attention maps in different heads indicate two points: **1)** The kernels for local operation have a global perceptive field with long-distance information. **2)** Different heads pay attention to different reference parts, *i.e.* relevant background reference, foreground content, overall tone, etc.

5. Conclusion

This paper proposes an effective network GKNet to learn global-aware harmony kernels for image harmonization, including harmony kernel prediction and harmony kernel modulation branches. For harmony kernel prediction, we propose LRE to extract long-term references and KPB to predict global-aware kernels. To better fuse long-term context, we design SCF to select relevant references. For harmony kernel modulation, we employ the predicted kernels for harmonization with location awareness. Extensive experiments demonstrate that our proposed algorithm outperforms the state-of-the-art algorithms on the iHarmony4 dataset and real image composition datasets.

Acknowledgements. This work was supported by a Grant from The National Natural Science Foundation of China (No. 2021YFB2012300).

References

[1] Steve Bako, Thijs Vogels, Brian McWilliams, Mark Meyer, Jan Novák, Alex Harvill, Pradeep Sen, Tony Derose, and Fabrice Rousselle. Kernel-predicting convolutional networks for denoising monte carlo renderings. *ACM Trans. Graph.*, 36(4):97–1, 2017. 3, 4

[2] Tim Brooks and Jonathan T Barron. Learning to synthesize motion blur. In *CVPR*, pages 6840–6848, 2019. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 4

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 4

[5] Xuhai Chen, Jiangning Zhang, Chao Xu, Yabiao Wang, Chengjie Wang, and Yong Liu. Better" cmos" produces clearer images: Learning space-variant blur estimation for blind image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1651–1661, 2023. 4

[6] 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. 7

[7] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, June 2020. 1, 2, 3, 5, 6, 7, 8

[8] Xiaodong Cun and Chi-Man Pun. Improving the Harmony of the Composite Image by Spatial-Separated Attention Module. *IEEE TIP*, 29:4759–4771, 2020. 1, 3, 4, 6

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[10] Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *WACV*, pages 3560–3569, 2021. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 4

[12] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 4

[13] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *CVPR*, pages 10571–10580, 2021. 3

[14] Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, Yang Liu, and Jianjun Zhao. Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining. In *AAAI*, volume 35, pages 1487–1495, 2021. 3, 4

[15] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *ICCV*, pages 14870–14879, October 2021. 3, 4, 6, 7

[16] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *CVPR*, pages 16367–16376, June 2021. 3, 6, 7

[17] Yucheng Hang, Bin Xia, Wenming Yang, and Qingmin Liao. Scs-co: Self-consistent style contrastive learning for image harmonization. In *CVPR*, pages 19710–19719, June 2022. 2, 3, 6, 7

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[19] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 4

[20] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Transactions on graphics (TOG)*, 25(3):631–637, 2006. 2

[21] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 3

[22] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two transformers can make one strong gan. *arXiv preprint arXiv:2102.07074*, 1(3), 2021. 4

[23] Younghyun Jo, Seoung Wug Oh, Jaeyeon Kang, and Seon Joo Kim. Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation. In *CVPR*, pages 3224–3232, 2018. 3

[24] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson W.H. Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *ECCV*, 2022. 3, 6, 7

[25] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[26] Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, and Song Wang. Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting. In *CVPR*, pages 1869–1878, 2022. 3

[27] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *CVPR*, pages 510–519, 2019. 4

[28] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[29] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. Region-aware adaptive instance normalization for image harmonization. In *CVPR*, pages 9361–9370, June 2021. 1, 2, 3, 4, 6, 7

[30] Rui Liu, Hanming Deng, Yangyi Huang, Xiaoyu Shi, Lewei Lu, Wenxiu Sun, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fuseformer: Fusing fine-grained information in

transformers for video inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14040–14049, 2021. 4

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 4

[32] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *CVPR*, pages 2502–2510, 2018. 3, 4

[33] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *CVPR*, pages 670–679, 2017. 3

[34] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. In *ACM SIGGRAPH 2003 Papers*, pages 313–318. ACM New York, NY, USA, 2003. 2

[35] François Pitié and Anil Kokaram. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. *IPSJ Transactions on Computer Vision and Applications*, 2007. 2

[36] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 2

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3, 4

[38] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. Foreground-aware semantic representations for image harmonization. In *WACV*, pages 1620–1629, January 2021. 2, 3, 4, 5, 6, 7

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

[40] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. In *European Conference on Computer Vision*, pages 31–44. Springer, 2010. 2

[41] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017. 1, 2, 3, 6, 7

[42] Thijs Vogels, Fabrice Rousselle, Brian McWilliams, Gerhard Röthlin, Alex Harvill, David Adler, Mark Meyer, and Jan Novák. Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics (TOG)*, 37(4):1–15, 2018. 3

[43] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *ICCV*, pages 4692–4701, 2021. 4

[44] 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. *arXiv preprint arXiv:2207.04788*, 2022. 3, 6

[45] Su Xue, Aseem Agarwala, Julie Dorsey, and Holly Rushmeier. Understanding and improving the realism of image composites. *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012. 2

[46] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[47] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. In *CVPR*, pages 2736–2746, 2022. 4

[48] Jiangning Zhang, Xiangtai Li, Jian Li, Liang Liu, Zhucun Xue, Boshen Zhang, Zhengkai Jiang, Tianxin Huang, Yabiao Wang, and Chengjie Wang. Rethinking mobile block for efficient neural models. *arXiv preprint arXiv:2301.01146*, 2023. 4

[49] Jiangning Zhang, Xiangtai Li, Yabiao Wang, Chengjie Wang, Yibo Yang, Yong Liu, and Dacheng Tao. Eatformer: improving vision transformer inspired by evolutionary algorithm. *arXiv preprint arXiv:2206.09325*, 2022. 4

[50] Jiangning Zhang, Chao Xu, Jian Li, Wenzhou Chen, Yabiao Wang, Ying Tai, Shuo Chen, Chengjie Wang, Feiyue Huang, and Yong Liu. Analogous to evolutionary algorithm: Designing a unified sequence model. *Advances in Neural Information Processing Systems*, 34:26674–26688, 2021. 4

[51] Jiangning Zhang, Chao Xu, Jian Li, Yue Han, Yabiao Wang, Ying Tai, and Yong Liu. Scsnet: an efficient paradigm for learning simultaneously image colorization and super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3271–3279, 2022. 4

[52] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *ECCV*, pages 269–284, 2018. 4