Spatial-Separated Curve Rendering Network for Efficient and High-Resolution Image Harmonization

Jingtang Liang[1*], Xiaodong Cun[2*], Chi-Man Pun[1✉], and Jue Wang[2]

[1] University of Macau, Macau, China
[2] Tencent AI Lab, Shenzhen, China
mb95464@connect.umac.mo, cmpun@umac.mo, {vinthony, arphid}@gmail.com

Abstract. Image harmonization aims to modify the color of the composited region according to the specific background. Previous works model this task as a pixel-wise image translation using UNet family structures. However, the model size and computational cost limit the ability of their models on edge devices and higher-resolution images. In this paper, we propose spatial-separated curve rendering network (S^2CRNet), a novel framework to prove that the simple global editing can effectively address this task as well as the challenge of high-resolution image harmonization for the first time. In S^2CRNet, we design a curve rendering module (CRM) using spatial-specific knowledge to generate the parameters of the piece-wise curve mapping in the foreground region and we can directly render the original high-resolution images using the learned color curve. Besides, we also make two extensions of the proposed framework via cascaded refinement and semantic guidance. Experiments show that the proposed method reduces more than 90% parameters compared with previous methods but still achieves the state-of-the-art performance on 3 benchmark datasets. Moreover, our method can work smoothly on higher resolution images with much lower GPU computational resources. The source codes are available at: http://github.com/stefanLeong/S2CRNet.

1 Introduction

Image composition (or image splicing in multimedia security) is a popular and necessary tool for image editing. However, in addition to the serrated edges caused by the irregular borders, the "style" disharmony occurs when we directly copy source regions (foreground) to the host image (background). The disharmony will degrade the quality of the composited images, which also can be distinguished by the human eyes easily. In general, handling this gap requires the professional editing of the well-knowledged experts. Thus, the task of image

* These authors contribute equally to this work.
✉ Corresponding author

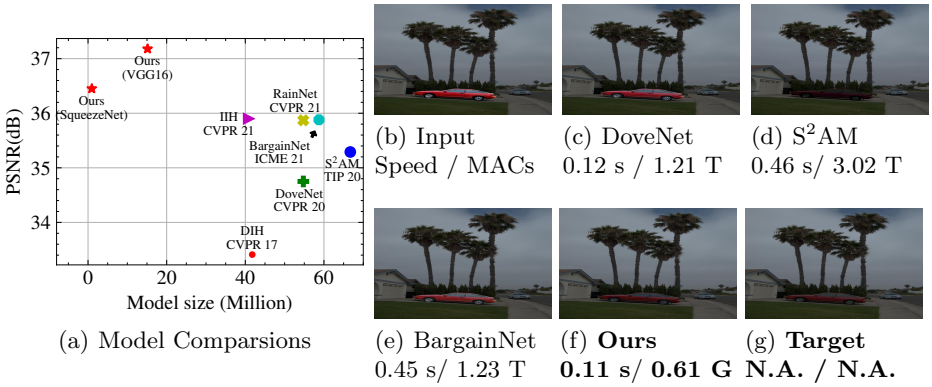
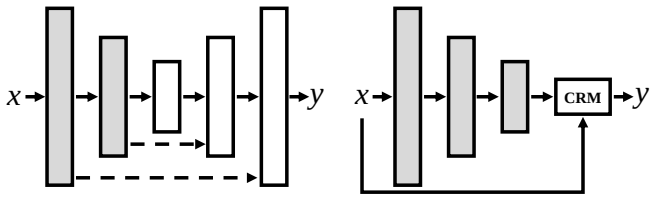


Fig. 1: (a) Our methods outperform other methods using much less parameters under the same setting (testing in 256×256 resolution). (b)-(f) Given a high-resolution image (originally 2048×2048 in this example), our method shows much better performance, lower computational cost (MACs) and faster speed than previous methods.

harmonization aims to squeeze this gap by leveraging some advanced algorithms, which also has a broad impact on image editing, relighting and augmented reality [38,22].

Traditional image harmonization methods intend to manually adjust and modify the specific features in the composite images, such as color [21,29], illumination [35] and texture [33], *etc.*. However, the hand-crafted and statistic low-level features cannot work well for the diverse composite images in complicated real world. Since the deep convolutional neural network (CNN) has reached impressive performance in many computer vision tasks, several attempts have also been made to address image harmonization tasks. For example, the semantic clues [34,32], the spatial differences of the neural network [5,12] and generative adversarial network (GAN [9]) based methods [4,3] have been proposed following the encoder-decoder based structures (UNet [30,18]) for pixel-wise prediction. Thus, as shown in Figure 1(a), the speed and computational cost are sensitive to image resolution because those structures require to predict the pixel-wise results. Besides, their model sizes are too large for the edge devices, such as mobile phone. The problems mentioned above restrict the applying range of their methods since the real-world images editing are at any resolution. Furthermore, further evaluations at high-resolution images would be also downgraded from these inefficiencies.

Differently, in this paper, we rethink the image harmonization in a totally different way: Reviewing the image harmonization process in image editing software (*e.g.* PhotoShop), experts tend to adjust the global properties (color curve, illuminant, *etc.*) over the whole images rather than the pixel-wise color adjustment. Thus, the global editing can be enabled by considering those properties as



(a) Previous methods (b) Ours

Fig. 2: We learn global mappings for image harmonization and are totally different from previous methods [5,4,3,32,12,34,11,23] that consider it as a pixel-wise image-to-image translation task.

the mapping function of the pixels intensities. Moreover, this global adjustment is reliably efficient at any resolution images without extra expense of computational cost.

Above observation inspires us to doubt the effect of the locally-aware editing networks in previous image harmonization methods and learning the *global editing curves* of the composite foreground in terms of efficiency as shown in Figure 2. Hence, a novel curve rendering module (CRM) is designed to produce the image-adaptive parameters of the curves that we will use to render the composite image. Specifically, we first separate the composite image into foreground/background regions using the given foreground mask. Then, we extract the global high-level features from both regions by a shared pre-trained general feature extractor (SqueezeNet [17] or VGG16 [31]). Particularly in CRM, the extracted features from foreground / background are learnt by a single layer linear projection for each region separately. Finally, the combination of these two spatial-specific features will be represented as the parameters of color curves, and we render the original foreground for each color channel with the approximate curves we learned.

Furthermore, we also make two extensions to the proposed framework. On one hand, we propose *semantic*-CRM. Since different foregrounds represent different categories, we learn the class-aware feature embeddings for each category individually by the user-guided foreground semantic encoding. On the other hand, we propose the *cascaded*-CRM, which is also inspired by the photo editing software since the image editing process generally contains multiple steps. In our implementation, we predict different domain embedding to achieve this goal via a cascaded prediction. Benefit by the proposed framework, our method shows a significantly better performance than previous state-of-the-art image harmonization networks with only 2% (25% using VGG16 backbone) of the parameters. Besides, our method can also run much faster than most previous methods with few computation cost on high-resolution images.

Our main contributions are summarized as follows:

- We find that the image harmonization can work well with the global editing method for the first time. To this end, we introduce a novel spatial-separated curve rendering network (S^2CRNet), which also enables our method for efficient and high-resolution image harmonization.
- We show the extension ability of the proposed S^2CRNet via better backbones or enhanced curve rendering module (CRM) via the Cascaded-CRM and Semantic-CRM.
- Experiments show that our method can achieve state-of-the-art performance and run much faster than the previous methods, while using fewer parameters and lower computational cost.

2 Related Works

Image Harmonization. Traditional image harmonization methods aim at improving composite images via low-level appearance features, such as manually adjusting global color distributions [2,1], applying gradient domain composition [19,28] or manipulating multi-scale transformation and statistical analysis [33]. Although these methods achieve preliminary results in harmonization tasks, the realism of the composite images cannot be visually guaranteed.

As the deep learning approaches has been successfully applied to the computer vision tasks, [39] back-propagate a pre-trained visual discriminator model to change the appearance harmony of the composite images. Later, further researches consider this task as an image to image translation problem. For example, additional semantic decoder [34] and pre-trained semantic feature [32] are used to ensure the semantic consistence between the composite inputs and harmonized outputs. Another noticeable idea is to model the differences between the foreground and background with the given mask. For example, novel spatial-separated attention module [5,12] under image-to-image translation framework; Domain-guided features as the discriminator of GAN [4] and as additional input [3]; masked-guided spatial normalizations [23,11] for the foreground and background respectively. However, all the previous deep networks still model this task as a pixel-wise image to image translation problem using an encoder-decoder structure, which suffers from computational inefficiency and may degrade the performance and visual quality on high-resolution inputs.

Efficient Network Design for Image Enhancement. Efficient networks designed for edge devices have also been widely-discussed in computer vision tasks [15]. For image enhancement, [8] introduce the deep bilateral learning for high-resolution and real-time image processing on mobile devices. Also, learning the image-adaptive global style features shows promising results in Exposure [16] , CURL [26] and 3DLUT [36] for global image enhancement. Besides, Guo *etal*. [10] design a high-order pixel-wise curve function for low-light enhancement. Since our image harmonization task can be considered as a *regional* image enhancement problem, it is natural to leverage the style curve to image harmonization tasks. However, different from the networks for image enhancement [16,26,36] and low-light enhancement [10], image harmonization methods

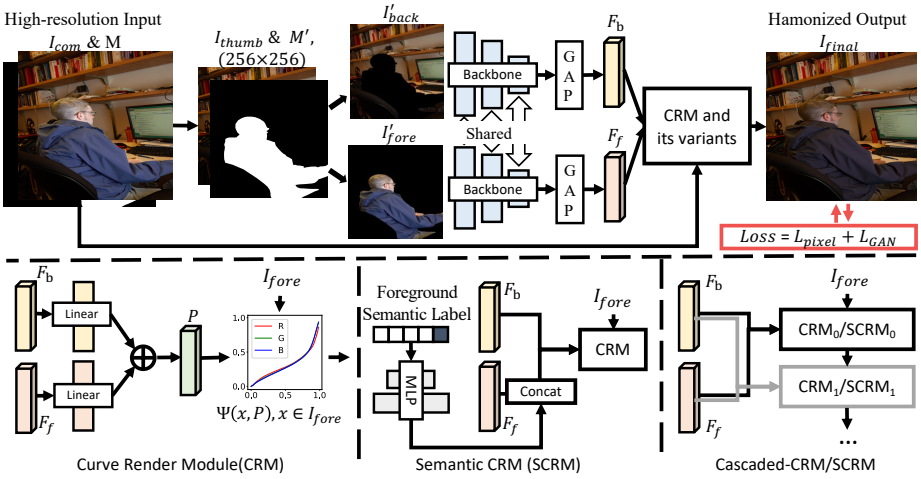


Fig. 3: The overview of the S^{2}CRNet, including CRM and its two variants: SCRM and Cascaded-CRM/SCRM.

rely on regional modification under the guidance of the background. Thus, we design the network structure and learn global mapping functions on this task for the first time.

3 Method

We first show the overall network structure of the proposed method. Then, we give the details of Curve Rendering Module (CRM) and its variants, which are the key components in S^{2}CRNet, including CRM, *Semantic*-CRM (SCRM) and their cascaded extensions. Finally, we discuss the loss functions.

3.1 Overall Network Structure

As shown in Figure 3, given a high-resolution composite image $I_{com} \in \mathbb{R}^{3\times H\times W}$ and its binary mask $M \in \mathbb{R}^{1\times H\times W}$ of the corresponding foreground, we first get the thumbnail image $I_{thumb} \in \mathbb{R}^{3\times h\times w}$ and the mask $M' \in \mathbb{R}^{1\times h\times w}$ by down-sampling I_{com} and M with a factor of H/h for fast inference and the minimal computational cost. For the spatial-separated feature encoding, we first segment the thumbnail image I_{thumb} into foreground and background via the mask M' and inverse mask $M'_{inv} = 1 - M'$, respectively. Next, given the foreground $I_{fore} = I_{thumb} \times M'$ and background $I_{back} = I_{thumb} \times M'_{inv}$ images, we use a shared domain encoder Φ to extract the spatial-separated features for foreground and background respectively. Here, we choose the SqueezeNet [17] as the domain encoder (backbone), which is pre-trained on the ImageNet [7] and we only use the first 12 layers to get deeper color feature embedding. We

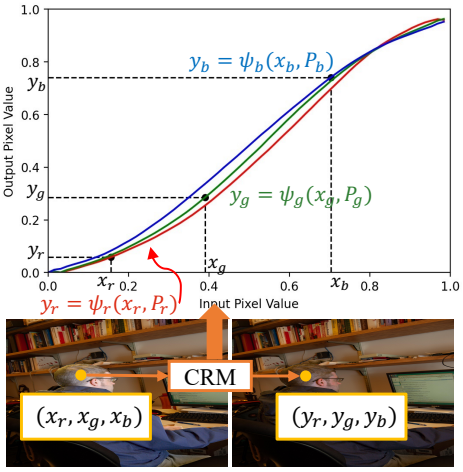


Fig. 4: CRM maps the input pixels to the target pixels using curve function $\psi(\cdot)$, where the parameters P of $\psi(\cdot)$ are learnt from the embeddings of the spatial-aware encoders.

also try different backbones (*e.g.*, VGGNet [31]) to achieve better performance as shown in Table 1. While considering the purpose of this paper is for efficient and high-resolution image harmonization, thus we use SqueezeNet as our default backbone for its good balance between the efficiency and effectiveness.

After obtaining the embedding foreground $F_{fore} \in \mathbb{R}^{D\times h'\times w'}$ and background $F_{back} \in \mathbb{R}^{D\times h'\times w'}$ features from the domain encoder, we squeeze the foreground/background feature dimensionally via the global average pooling to avoid the influence of spatial information. Then, foreground $F_f \in \mathbb{R}^D$ and background $F_b \in \mathbb{R}^D$ are learnt to generate the parameters of the color curve and render the channel-wise color curve via the proposed *Curve Rendering Module* automatically. We will discuss the details and its variants in the later sections.

3.2 Curve Rendering Modules and its Variants

We first introduce the basic idea behind the proposed network via the *Curve Rendering Module* (CRM). Then, we discuss two different extensions using the semantic label and recurrent refinement.

Curve Rendering Module (CRM). Most previous image harmonization methods [4,5,32,12] consider this task as a pixel-wise image to image translation task, which is heavy and only works on certain resolution as we discussed in the related works. Differently, we model this task as a global region image enhancement task. Thus, our goal of CRM is try to adjust the foreground color under the given background.

To achieve the above goal, as shown in Figure 3, after obtaining the spatial-separated foreground embedding F_f and background embedding F_b from the

domain encoder separately, we first embed these spatial-aware features using two projection functions $\phi_f(\cdot)/\phi_b(\cdot)$ for foreground/background correspondingly, where each projection function is a single linear layer with ReLU activation. Then, to harmonize the foreground under the guidance of the related background features, we get $P \in \mathbb{R}^{3L}$ by performing channel-wise addition between $\phi_f(F_f)$ and $\phi_b(F_b)$. Here, L includes the parameters of R, G, B color channels and each channel has $L = 64$ piece parameters for the balance between the computational complexity and performance.

Since this hybrid feature P contains both the information from the background and foreground, it can be a good representation for the guidance of the foreground editing. To better modeling the color-wise changes, we consider the mappings between intensities rather than the semantic information. Thus, we choose the color curve as the editing tool and make it differentiable [16] by approximating L levels monotonous piece-wise linear function, and then rendering the original pixels in the foreground region. As shown in Figure 4, for each pixels (x_r, x_g, x_b) in the foreground of the original composited image, we use CRM to map it with the learnt color curve. Here, the mappings of each intensity is identical and not related to the specific location or semantic information. The parameters of the piece-wise linear function is provided and learnt by the spatial-separated encoder and each channel is learnt individually.

Mathematically, after getting the mixed embedding for each channel $P^c = [p_0, p_1, p_2, \ldots, p_{L-1}]$, we render the foreground I^c_{fore} $(c \in \{R, G, B\})$ of the composite image via the curve rendering function $\psi(I^c_{fore}, P^c)$, which can be denoted as:

$$\psi_c(I^c_{fore}, P_c) = \frac{1}{\sum_{j=0}^{L-1} p_j} \sum_{i=0}^{L-1} p_i \xi\left(x - \frac{i}{L}\right), x \in I^c_{fore},$$

$$\text{where} \quad \xi(y) = \begin{cases} 0, & y < 0 \\ y, & 0 \le y < \frac{1}{L} \\ 1, & y > \frac{1}{L} \end{cases} \tag{1}$$

Finally, the harmonized image can be obtained by the combination of the original background: $I_{final} = \Psi(I_{fore}, P) + I_{back}$.

Semantic CRM. Previous methods [4,5] intend to obtain a unified harmonization model for any foreground images without any specific semantic knowledge. However, the semantic information is also important for the image harmonization [34,32] and it does not make sense if we apply the same style to harmonize different categories (*e.g.* Car and $Person$). Since we have supposed that the linear layers in the CRM contain the domain knowledge of the foreground, we make a further step by adding extra semantic label of the foreground object to our vanilla CRM.

As shown in Figure 3, given the semantic label d of the foreground region, we first embed the labels using a two-layer Multi-layer Perceptron (MLP), obtaining the semantic-aware embedding D. Then, we concatenate the embedded feature from the network Φ and the label embedding D to the CRM. For semantic label definition, we analyze the categories of the foreground regions in iHarmony4 and

divide it into 5 classes as guidance, including *Person*, *Vehicle*, *Animal*, *Food* and others. More details can be found in the supplementary materials.

Cascaded CRM/SCRM. It is natural for the image editing tools to adjust the images with multiple steps for better visual quality. Inspired by this phenomenon, we extend our CRM (or SCRM) via the cascaded refinements. To reduce the inference time and learn a compact model, we keep the global features from the backbone unchanged and generate multi-stage heads and give the supervisions of each stage.

As shown in Figure 3, given the global foreground features F_f and background features F_b from the backbone, we firstly generate P_0 via a CRM and get its rendered image I_0 using $\Psi_c(I^c_{fore}, P_0)$. Then, we use another set of linear layers to predict the parameters P_n from the same global features (F_f, F_b) and rendering the curve using the previous prediction I_{n-1} via $\Psi_c(I_{n-1}, P_n)$. We set n equals to 2 to ensure the high efficiency as well as the high harmonization quality.

3.3 Loss Function

We consider image harmonization as a supervised problem. Specifically, we measure the difference between the target and the corresponding rendered images (for each stage) in the composited region. Thus, we use relative L_1 loss between the predicted foreground and the target via the foreground mask M. Besides, for better visual quality, we also leverage the adversarial loss [9] to our framework. We give the details of each part as follows.

Relative L_1 Loss L_{pixel}. Another key idea to make our method work is that we only calculate the metric between the foreground of the predicted image and the target, where the differences are only measured in a single domain. Thus, inspired by recent works in watermark removal [14,6], we perform the pixel-wise L_1 loss in the foreground region M by masking out the background pixels and setting the meaningful region. Specifically, giving the rendered images I_n in each stage, we calculate the loss over the masked region:

$$L_{pixel} = \sum_{n=1}^{N} \frac{||M \times I_n - M \times I_{gt}||_1}{sum(M)} \tag{2}$$

where $N = 2$ is the number of iterations.

Adversarial Loss L_{adv}. By considering the proposed S^{2}CRNet as the generator G, we also utilize an additional discriminator D to identify the naturalness of the color. In detail, we use a standard 5 layers CONV-BN-RELU discriminator [40] and leverage a least squares GAN [25] as criteria. Then, the generator is learnt to fool the discriminator and the discriminator is trained to identify the real or fake feed images iteratively.

Overall, our algorithm can be trained in an end-to-end function via the combination of the losses above: $L_{all} = \lambda_{pixel} L_{pixel} + \lambda_{adv} L_{adv}$, where all the hyper-parameters (λ_{pixel} and λ_{adv}) are empirically set to 1 for all our experiments.

(a) (b) (c) (d) (e) (f) (g)

Fig. 5: Comparisons with other methods on iHarmony4 Dataset. From left to right are (a)Input (b) DoveNet [4], (c) BarGainNet [3], (d) S^{2}AM [5], (e) S^{2}CRNet-S (Ours), (f) S^{2}CRNet-V (Ours) and (g) Target. Here, we visualize the input mask as yellow for easy reading. S^{2}CRNet-S and S^{2}CRNet-V denote our method employs SqueezedNet and VGG16 backbone, respectively.

4 Experiments

4.1 Implementation Details

We implement our method in Pytorch [27] and train on a single NVIDIA TITAN V GPU with 12GB memory. The batch size is set to 8 and we train 20 epochs (50 epochs for VGG16 backbone) for convergence. All the images are resized to 256×256 and random cropped and flipped for fair training and evaluation as previous methods [4,5]. We leverage the AdamW optimizer [24] with the learning rate of 2×10^{-4}, the weight decay value of 10^{-2} and momentum of 0.9.

As for evaluation, we validate our approaches on the iHarmony4 using Mean-Square-Errors (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural SIMilarity (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) as criteria metrics. Since DIH99 does not contain the target images, we conduct the subjective experiments.

4.2 Comparison with Existing Methods

Performance Comparison on iHarmony4. We compare our methods with other state-of-the-art image harmonization algorithms, including DoveNet, S^{2}AM, BargainNet, IIH [11], RainNet [23], *etc.*. In our experiments, we choose the

Table 1: Comparisons on iHarmony4. The best and the second best are marked as boldface and underline respectively.

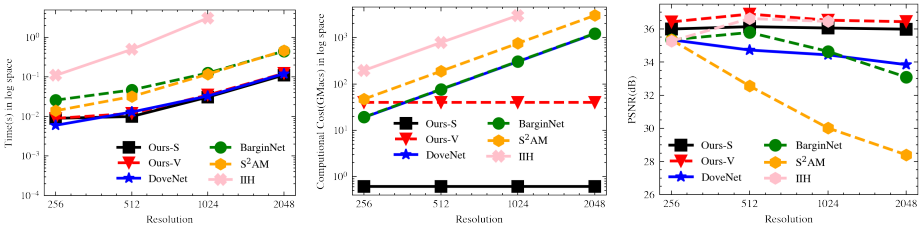
Sub-dataset		HCOCO		HAdobe5k		HFlickr		Hday2night		All	
Evaluation metric	# Param.	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑
Input Composition	-	67.89	34.07	342.27	28.14	260.98	28.35	107.95	34.01	170.25	31.70
Lalonde & Efros [21]	-	110.10	31.14	158.90	29.66	329.87	26.43	199.93	29.80	150.53	30.16
Xue *etal.* [35]	-	77.04	33.32	274.15	28.79	249.54	28.32	190.51	31.24	155.87	31.40
Zhu *etal.* [39]	-	79.82	33.04	414.31	27.26	315.42	27.52	136.71	32.32	204.77	30.72
DIH[34]	41.76M	51.85	34.69	92.65	32.28	163.38	29.55	82.34	34.62	76.77	33.41
DoveNet[4]	54.76M	36.72	35.83	52.32	34.34	133.14	30.21	54.05	35.18	52.36	34.75
S^2AM[5]	66.70M	33.07	36.09	48.22	35.34	124.53	31.00	48.78	35.60	48.00	35.29
BargainNet[3]	58.74M	24.84	37.03	39.94	35.34	**97.32**	31.34	50.98	35.67	37.82	35.88
IIH[11]	40.86M	24.92	37.16	43.02	35.20	105.13	31.34	55.53	35.96	38.71	35.90
RainNet[23]	54.75M	31.12	36.59	42.84	36.20	117.59	31.33	**47.24**	36.12	44.50	35.88
S^2CRNet-SqueezeNet	**0.95M**	28.25	37.65	44.52	35.93	115.46	31.63	53.33	36.28	43.20	36.45
S^2CRNet-VGG16	15.14M	**23.22**	**38.48**	**34.91**	**36.42**	98.73	**32.48**	51.67	**36.81**	**35.58**	**37.18**

Cascaded-SCRM model in different backbones (SqueezeNet and VGG16 as shown in Table 1), where the semantic labels are generated by a pre-trained segmentation model [37]). All previous methods are tested using their official implementations and pre-trained models for fair comparison. As shown in Table 1, even training and testing on 256×256 limits the high-resolution performance, our S^2CRNet-SqueezeNet only use 2% of the parameters to achieve the state-of-the-art performance in PSNR metric, which demonstrates the effectiveness of the proposed network. On the other hand, when using VGG16 backbone (S^2CRNet-VGG16), our method outperforms other related methods by a clear margin and still uses only 40% of the parameters. Moreover, the proposed method also works better even on higher-resolution images, which will be discussed in later section. Besides the numeric comparison, our proposed method also obtains better visual quality than others. Qualitative examples in Figure 5 show that the proposed method can generate harmonized results that are more realistic than other methods, which further indicates the benefits of the proposed framework. More visual comparisons are presented in the supplementary materials.

Performance on real-world composite datasets. Since the real-wold image composition is still different from the synthesized dataset, we evaluate the proposed method (S^2CRNet-SqueezeNet) and existing methods (DIH, DoveNet, BarginNet) by subjective experiments on DIH99. In detail, we randomly shuffle the displaying order of all the images and invite 18 users to select the most realistic results. As shown in Table 2, the proposed method gets the most votes with faster inference time and fewer model parameters as discussed previously.

Table 2: User study on DIH99 [34] test set.

Method	Input	DIH	DoveNet	BargainNet	Ours
Total votes	224	385	403	328	**442**
Preference	12.57%	21.60%	22.62%	18.41%	**27.80%**



(a) Running time. (b) Computational cost. (c) Performance.

Fig. 6: The influence of the image resolution from different aspects. Note that all experimental values are transformed into *log* scale. IIH [11] cause out of memory error on 2048 × 2048 images.

Additionally, we evaluate our method on RealHM dataset following [20] and summarize the qualitative results in Table 3. From Table 3, our method outperforms others at SSIM and LPIPS metrics with much less parameters and processing time. Particularly, we obtain the similar performance compared with SSH [20] at PSNR and MSE metrics while SSH is trained on 2× larger dataset with 10× lager model and stronger data augmentation. The harmonization results on DIH99 and RealHM dataset effectively demonstrate that our method has good generalization ability on the real-world applications. More details of the user study and more harmonization results of the real composite samples are shown in the supplementary.

High-Resolution Image Harmonization. We conduct extra experiments on the HAdobe5k sub-dataset in iHarmony4 to verify the speed and performance of the proposed method on higher-resolution. As experiment setup, we resize the source composite images with the square resolution of 256, 512, 1024 and 2048, and test the average processing time, computational cost and PSNR scores on the same hardware platform. Since other state-of-the-art methods (DoveNet, BargainNet, S^{2}AM, IIH) employ the fully convolutional encoder-decoder structures, they can be tested directly in higher resolution. As for our method, we test two backbones of the proposed S^{2}CRNet and donate them as Ours-S (SqueezeNet as backbone) and Ours-V (VGG16 as backbone) shown in Figure 6.

As shown in Figure 6(a), we plot the speed of different image harmonization methods in the *log* space. All the methods suffer a speed degradation with the

Table 3: Quantitative comparisons on RealHM dataset [20].

	PSNR↑	MSE↓	SSIM↑	LPIPS↓	Time↓	Parameters↓
DoveNet	27.41	214.11	94.14	0.049	0.081s	54.76M
S^{2}AM	26.77	283.27	93.66	0.096	0.282s	66.70M
SSH	**27.91**	**206.85**	94.79	0.039	0.153s	15.19M
Ours	27.89	229.64	**96.16**	**0.025**	**0.012s**	**0.95M**

resolution increasing. However, our the research quality code of Ours-S and Ours-V runs much faster (0.1s around on a 2048×2048 image) than all other methods and is nearly $5\times$ faster than S^2AM and BargainNet. Also, since we use a fixed size input, the required computation cost of our method still much less than previous methods as the resolution increasing as shown in Figure 6(b). In terms of the harmonization quality, there are also some interesting phenomenons. As shown in Figure 6(c), most of other methods confront a significant performance decline as the resolution increases. It might be because the encoder-decoder based structure will produce different reception fields of original images and then downgrade its performance. Differently, our methods maintain the higher performance at all resolutions.

4.3 Ablation Studies

We conduct the ablation experiments to demonstrate the effectiveness of each component in the proposed S^2CRNet. All the experiments are performed on both HCOCO and iHarmony4 with same configurations using the SqueezeNet backbone. More ablation studies are presented in the Appendix.

Loss Function. As shown in Table 4 Model A to C, we compare the performance using different loss functions. Since background and foreground domains are different, restricting the loss function on the masked region by using relative L_1 (rL_1) rather than L_1 loss helps a lot. Besides, L_{adv} are used to improve the realism of the predicted result.

Encoder Design Φ. Extracting and learning the global foreground and background features individually (Ours in Model C in Table 4) are also the keys to facilitate the performance of the whole framework. As shown in Table 4 and Figure 7, compared with other alternatives that extract the global features using the foreground region only (I_{fore} in Model D) and the full image (I_{com} in Model E), spatial-separated encoder shows a much better performance due to domain separation.

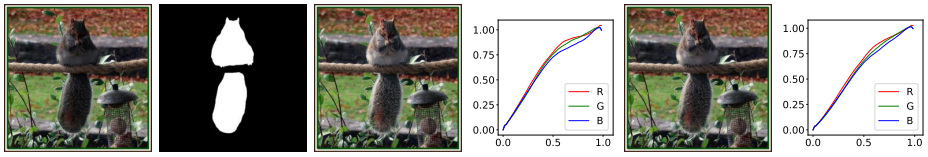
CRMs. The numerical metrics of different CRMs have been listed in Table 4, both *Cascaded*-CRM (Model F) and *Cascaded*-SCRM (Model G) hugely

Table 4: Ablation studies.

#	Loss		Network		HCOCO		iHarmony	
	L_{pixel}	L_{adv}	Φ	CRM	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑
-	Original Input				67.89	34.07	170.25	31.70
A	L_1		Ours	✓	67.64	34.08	114.65	32.11
B	rL_1		Ours	✓	28.43	37.59	46.79	36.20
C	rL_1	✓	Ours	✓	29.45	37.51	45.17	36.27
D	rL_1	✓	I_{fore}	✓	34.62	36.98	79.73	34.57
E	rL_1	✓	I_{com}	✓	58.53	34.69	88.61	33.88
F	rL_1	✓	Ours	C	28.47	37.60	44.08	36.41
G	rL_1	✓	Ours	CS	**27.40**	**37.72**	**43.20**	**36.45**

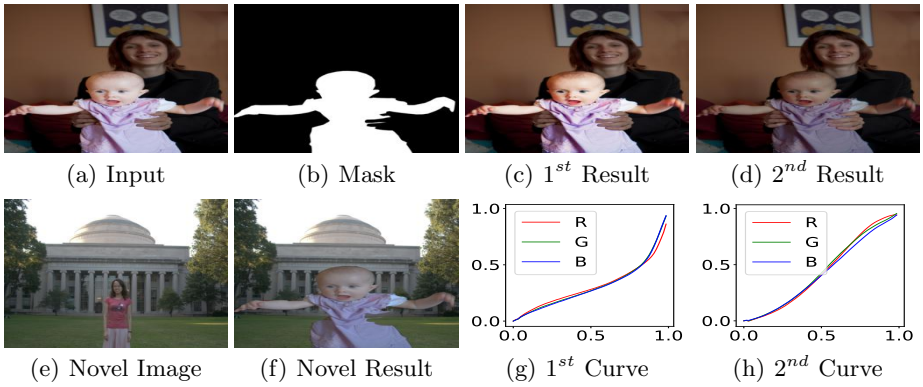
(e) Input (f) I_{fore} (g) I_{com} (h) Ours

Fig. 7: The influence of different encoder designs.



(a) Input (b) Mask (c) *Animal* (d) *Animal* (e) *Person* (f) *Person*

Fig. 8: Results and rendering curves of SCRM using different foreground semantic labels (*Person*, *Animal*).



(a) Input (b) Mask (c) 1^{st} Result (d) 2^{nd} Result

(e) Novel Image (f) Novel Result (g) 1^{st} Curve (h) 2^{nd} Curve

Fig. 9: Given a composite image (a) and its mask (b), Cascaded-CRM learns to generate different harmonization results (c, d) via curves (g, h). Also,our method can harmonize the current foreground via novel backgrounds (e, f).

improve the base model (Model E). To further explore the influence of each variants, firstly, we show the importance of semantic labels. As shown in Figure 8, different semantic labels will produce different style curves under the same input.

Then, for cascaded refinement, in Figure 9, cascaded refinement will produce different curves and achieve gradually better performance. Finally, the global color curves enable the proposed method to harmonize images with domain-aware features from novel images. In Figure 9, a novel background can also guide the harmonization of different foreground regions, which means our method can handle two objects harmonization by generating corresponding curves for each object individually.

5 Discussion and Real-World Application

By utilizing the global editing only, the proposed framework start a new direction for image harmonization which is efficient, flexible and transparent. As for efficiency, both performance and speed are better than previous methods. With respect to flexibility, images at any resolution can be edited without additional processing, like guided filter [34,13]. As for transparency, our method is a "white-box" algorithm because the learned curves can be further edited by the user to increase/decrease the harmony. On the other hand, since the search space of the global editing is bounded, the proposed method can be used directly for video harmonization without retraining on video datasets. We show a brief video harmonization result in the supplementary to demonstrate the potential of our framework.

6 Conclusion

In this paper, we investigate the possibility of global editing only for image harmonization for the first time. To this end, we present Spatial-separated Curve Rendering Network (S^2CRNet), a novel framework for efficient and high-resolution image harmonization. In detail, we utilize an efficient backbone to obtain spatial domain-aware features and the extracted features are used to generate the parameters of piece-wise curve function in the proposed curve render model. Besides, we extend the proposed curve rendering method to cascaded refinement and semantic-aware prediction. Finally, the learnt parameters are used to render the original high-resolution composite foreground. Experiments show the advantages of the proposed framework in terms of efficiency, accuracy and speed.

Acknowledgements. This work is supported in part by the University of Macau under Grant MYRG-2018-00035-FST and Grant MYRG-2019-00086-FST and in part by the Science and Technology Development Fund, Macau SAR, under Grant 0034/2019/AMJ, Grant 0087/2020/A2 and Grant 0049/2021/A.

References

1. Bychkovsky, V., Paris, S., Chan, E., Durand, F.: Learning photographic global tonal adjustment with a database of input/output image pairs. In: CVPR. pp. 97–104. IEEE (2011)
2. Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.Q.: Color harmonization. In: SIGGRAPH, pp. 624–630 (2006)
3. Cong, W., Niu, L., Zhang, J., Liang, J., Zhang, L.: BargainNet: Background-guided domain translation for image harmonization. In: ICME (2021)
4. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: CVPR. pp. 8394–8403 (2020)
5. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. TIP **29**, 4759–4771 (2020)
6. Cun, X., Pun, C.M.: Split then refine: Stacked attention-guided resunets for blind single image visible watermark removal. AAAI (2021)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255. Ieee (2009)
8. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. TOG **36**(4), 1–12 (2017)
9. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
10. Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: CVPR. pp. 1780–1789 (2020)
11. Guo, Z., Zheng, H., Jiang, Y., Gu, Z., Zheng, B.: Intrinsic image harmonization. In: CVPR. pp. 16367–16376 (June 2021)
12. Hao, G., Iizuka, S., Fukui, K.: Image harmonization with attention-based deep feature modulation. In: BMVC (2020)
13. He, K., Sun, J., Tang, X.: Guided image filtering. TPAMI **35**(6), 1397–1409 (2012)
14. Hertz, A., Fogel, S., Hanocka, R., Giryes, R., Cohen-Or, D.: Blind visual motif removal from a single image. In: CVPR. pp. 6858–6867 (2019)
15. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
16. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. TOG **37**(2), 1–17 (2018)
17. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. arXiv preprint arXiv:1602.07360 (2016)
18. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR. vol. 2017-Janua, pp. 5967–5976 (nov 2017). https://doi.org/10.1109/CVPR.2017.632, http://arxiv.org/abs/1611.07004
19. Jia, J., Sun, J., Tang, C.K., Shum, H.Y.: Drag-and-drop pasting. TOG **25**(3), 631–637 (2006)
20. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: Ssh: A self-supervised framework for image harmonization. In: ICCV. pp. 4832–4841 (2021)
21. Lalonde, J.F., Efros, A.A.: Using color compatibility for assessing image realism. In: ICCV. pp. 1–8. IEEE (2007)

22. Lee, D., Pfister, T., Yang, M.H.: Inserting videos into videos. In: CVPR. pp. 10061–10070 (2019)
23. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: CVPR. pp. 9361–9370 (June 2021)
24. Loshchilov, I., Hutter, F.: Fixing weight decay regularization in adam (2018)
25. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV. pp. 2794–2802 (2017)
26. Moran, S., McDonagh, S., Slabaugh, G.: Curl: Neural curve layers for global image enhancement. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 9796–9803. IEEE (2021)
27. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
28. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: SIGGRAPH, pp. 313–318 (2003)
29. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. CG&A **21**(5), 34–41 (2001)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
31. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
32. Sofiiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. In: WACV. pp. 1620–1629 (2021)
33. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. TOG **29**(4), 1–10 (2010)
34. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: CVPR (2017)
35. Xue, S., Agarwala, A., Dorsey, J., Rushmeier, H.: Understanding and improving the realism of image composites. TOG **31**(4), 1–10 (2012)
36. Zeng, H., Cai, J., Li, L., Cao, Z., Zhang, L.: Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. TPAMI (2020)
37. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV (2018)
38. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single portrait image relighting. In: ICCV (2019)
39. Zhu, J.Y., Krahenbuhl, P., Shechtman, E., Efros, A.A.: Learning a discriminative model for the perception of realism in composite images. In: ICCV. pp. 3943–3951 (2015)
40. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. ICCV (Mar 2017)