FRIH: Fine-Grained Region-Aware Image Harmonization

Jinlong Peng*, Zekun Luo, Liang Liu, Boshen Zhang

Tencent Youtu Lab
{jeromepeng, zekunluo, leoneliu, boshenzhang}@tencent.com

Abstract

Image harmonization aims to generate a more realistic appearance of foreground and background for a composite image. All the existing methods perform the same harmonization process for the whole foreground. However, the implanted foreground always contains different appearance patterns. Existing solutions ignore the difference of each color block and lose some specific details. Therefore, we propose a novel global-local two stages framework for Fine-grained Region-aware Image Harmonization (FRIH). In the first stage, the whole input foreground mask is used to make a global coarse-grained harmonization. In the second stage, we adaptively cluster the input foreground mask into several submasks. Each submask and the coarsely adjusted image are concatenated respectively and fed into a lightweight cascaded module, refining the global harmonization result. Moreover, we further design a fusion prediction module to generate the final result, utilizing the different degrees of harmonization results comprehensively. Without bells and whistles, our FRIH achieves a competitive performance on iHarmony4 dataset with a lightweight model.

Introduction

Image composition plays an essential role in image editing and generation (Liu et al. 2020; Azadi et al. 2020; Qiu et al. 2020; Cheng et al. 2020; Wang et al. 2020; Guo et al. 2019; Van den Oord et al. 2016). However, since the source of the implanted foreground object and the new background image are different, it is easy to cause an unrealistic perception of the composite image. Image harmonization is an important operation to address this issue, aiming to make the implanted foreground compatible with the background.

Traditional image harmonization methods mainly focused on low-level feature statistics, such as color distribution matching (Cohen-Or et al. 2006; Pitié and Kokaram 2007), gradient-domain compositing (Pérez, Gangnet, and Blake 2003; Jia et al. 2006; Tao, Johnson, and Paris 2013) and hybrid feature transferring (Sunkavalli et al. 2010). These methods limit the harmonization performance due to the lack of high level information. There are also several unsupervised or self-supervised image harmonization methods (Chen and Kae 2019; Zhan et al. 2020; Jiang et al. 2021), not relying on manual annotation. But they only work better in specific scenes, such as the portraits harmonization, lack of universality. Recently, the supervised encoder-decoder harmonization methods (Tsai et al. 2017; Cong et al. 2020; Ling et al. 2021) have achieved superior performance based on the construction of the large training datasets.

Figure 1: Case study on the coarse-grained harmonization problem of different foreground appearance patterns. On the top line, the implanted foreground is a person with green shirt and orange hair. The background contains more orange pixels, leading to a better harmony of the orange hair, while the saturation in the green shirt decreases in RainNet (Ling et al. 2021). On the bottom line, the implanted foreground is a gray sculpture and a red wreath. The background is brighter, resulting in a bright overall harmony process in RainNet (Ling et al. 2021), in which the red wreath looks good while the gray sculpture becomes too bright. Clearly, our FRIH could solve all the above problems well.

However, there is still a major problem of existing solutions. The implanted foreground always contains different appearance patterns. The similarity between the composite image and the real ground-truth image varies among different sub-regions. Some sub-regions in the composite image are relatively similar to the real ground-truth image, while others may be much far away from the target. Thus different sub-regions need different harmonization operations. Existing methods perform the same harmonization process for the whole foreground, ignoring the difference of each

*Corresponding author.

color block and losing some specific details. As shown in Figure 1, the foreground of the first case is a person with green shirt and orange hair. The foreground of the second case is a gray sculpture and a red wreath. If applying the coarse-grained harmonization method RainNet (Ling et al. 2021), only part of the foreground (the orange hair and the red wreath) has satisfactory performance, while the other part (the green shirt and the gray sculpture) is ineffective.

In order to solve the above problems, we propose a simple, novel and effective global-local framework for Fine-grained Region-aware Image Harmonization (FRIH), which is a two-stage network. The first stage includes the base network, i.e. a simple U-Net (Ronneberger, Fischer, and Brox 2015) alike network, in which the composite image and the whole input foreground mask are used to make a global coarse-grained harmonization. In the second stage, we adjust the global harmonization result according to the region-aware local feature. Specifically, we adaptively cluster the input foreground mask into several submasks by the corresponding pixel RGB values in the composite image. The number of the submasks is adaptive for different input images. Each submask and the coarsely adjusted image are concatenated respectively and fed into a lightweight cascaded module. Moreover, to utilize the different levels of the harmonization results comprehensively, we further design a fusion prediction module by fusing features from all the cascaded decoder layers together to generate the final result. Our two-stage network is trained end-to-end. If the two stages are trained separately, the submasks information can not affect the training of the first stage, not able to maximize the optimization of the overall performance.

Our FRIH method achieves a competitive performance in iHarmony4 (Cong et al. 2020) dataset with a lightweight model. The PSNR of our method is 38.19 dB. Specifically, the sub-regions details of the two cases in Figure 1 are both handled well by our method. Besides, our model has only 11.98 M parameters. The main contributions of this paper are as follows:

1. We propose a novel global-local framework for fine-grained region-aware image harmonization. The local submasks are generated adaptively to adjust the global coarse-grained harmonization result.

2. We design a lightweight cascaded module to integrate the global coarsely adjusted image and the region-aware local feature, refining the harmonization performance. And the fusion prediction module is further proposed to utilize the different degrees of harmonization results comprehensively.

3. Our FRIH achieves a competitive PSNR (38.19 dB) on iHarmony4 dataset. Besides, the parameters of our model are only 11.98 M.

Related Works

Statistics-based Harmonization Methods

Traditional image harmonization methods mainly focused on low-level feature statistics, such as color distribution matching (Pitie, Kokaram, and Dahyot 2005; Cohen-Or et al. 2006; Pitié and Kokaram 2007; Song et al. 2020), gradient-domain compositing (Pérez, Gangnet, and Blake 2003; Jia et al. 2006; Tao, Johnson, and Paris 2013) and hybrid feature transferring (Sunkavalli et al. 2010). These methods did not consider the realism of the composite images. Several other methods (Lalonde and Efros 2007; Xue et al. 2012; Zhu et al. 2015) further applied high-level image feature to design the visual reality assessment mechanism. Despite the better optimization standards, the basis of these methods is still statistics methods, limiting the harmonization performance. While our FRIH follows the currently mainstream supervised encoder-decoder framework, which has the potential for better harmonization performance.

Encoder-decoder Harmonization Methods

Comparing with traditional statistics-based methods, recent encoder-decoder harmonization methods have achieved superior performance. The pioneering end-to-end CNN method DIH (Tsai et al. 2017) encoded the input image and foreground mask, which was then decoded to the harmonized image and scene parsing image. Based on the encoder-decoder framework, several methods (Cun and Pun 2020; Hao et al. 2020; Cong et al. 2021) applied the attention mechanism to learn foreground and background appearance feature separately for harmonization. Moreover, RainNet (Ling et al. 2021) designed a region-aware adaptive instance normalization module to transfer the visual style from background to foreground. IIH (Guo et al. 2021b) proposed intrinsic image harmonization framework by disentangling the composite image into reflectance and illumination for further separate harmonization. Guo *et. al* (Guo et al. 2021a) integrated the transformer structure (Vaswani et al. 2017; Dosovitskiy et al. 2021) to the encoder-decoder harmonization network. Several recent methods (Ke et al. 2022; Liang et al. 2022) focused on high resolution image harmonization. However, all these methods only divided the foreground and background as different regions. They did not make a finer region division inside the foreground. Differently, our FRIH is the first method to propose the fine-grained region-aware harmonization framework, which could get more precise harmonization results.

Cascade Mechanism

Cascade mechanism is a direct but useful strategy in many computer vision tasks (Cai and Vasconcelos 2018; Chen et al. 2019). The key idea is to use an extra cascaded module to refine the results from the previous stages. Cascade R-CNN (Cai and Vasconcelos 2018) used a sequence of detectors trained with increasing IoU thresholds to gradually refine the detection results from the previous detectors. CU-Net (Liu et al. 2019) connected two U-Net networks and used two stage loss supervision for more accurate segmentation. Cascade EF-GAN (Wu et al. 2020) decomposed the expression editing task into three steps and applied three cascaded GANs to solve each of them. Most of these methods used the same architecture of base network to design their cascaded module, leading to a large increase of model size. Different from those methods, our cascaded module is lightweight and effective. The lightweight module together with the embedded fusion prediction module only accounts for 22.4% of the parameters of the whole FRIH network.

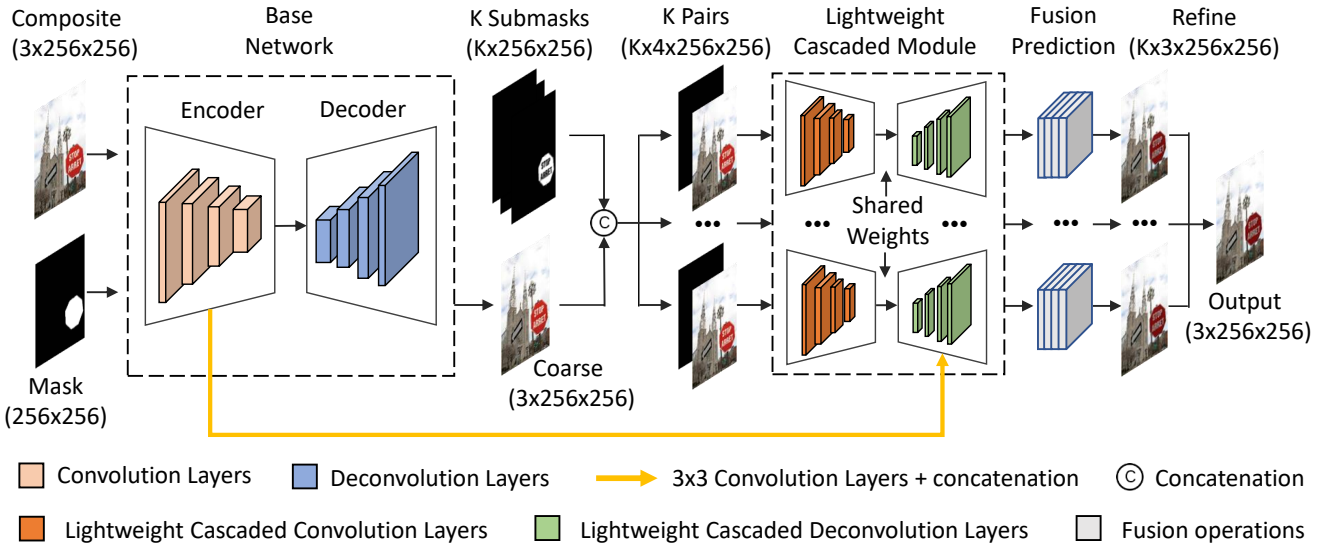

Figure 2: Our proposed FRIH framework. In the first stage, we feed the composite image and the foreground mask into the base network to obtain the coarsely adjusted image. In the second stage, the coarsely adjusted image and the extracted submasks are concatenated and fed into the lightweight cascaded module together with the embedded fusion prediction module, which adjusts the sub-regions and generates final refined harmonious images. Note that to keep the figure clean, we omit the skip connections between the encoder and the decoder, both in the base network and the lightweight cascaded module.

Proposed Method

Overview

The definition of image harmonization is to input a composite image I_c with the corresponding foreground mask and output the harmonious image $\hat{I}$. If there is a ground-truth image I of the composite image I_c, the optimization direction of the harmonization model is to make $\hat{I}$ close to I.

Our proposed fine-grained region-aware image harmonization framework is illustrated in Figure 2, which is a two-stage network. We use a simple U-Net (Ronneberger, Fischer, and Brox 2015) alike network as the base network in the first stage. In this stage, we feed the composite image and the foreground mask into the encoder-decoder network to obtain the global coarsely adjusted image. In the second stage, we first cluster the foreground mask into several submasks adaptively. Then, each submask and the coarsely adjusted image generated in the first stage are concatenated respectively and fed into the lightweight cascaded module, which makes full use of the region-aware local feature according to the submasks. Finally, we embed the fusion prediction module into the decoder of the lightweight cascaded module, to ultilize the different levels of the harmonization results comprehensively and generate the final harmonious images. The whole FRIH network is trained end-to-end.

Base Network

The base network inputs the composite image and the foreground mask, and outputs the global coarsely adjusted image. The base network is a simple U-Net (Ronneberger, Fischer, and Brox 2015) alike network, including an encoder and a decoder. The encoder has been downsampled for 7 times and the decoder has 7 deconvolution layers correspondingly. There is a skip connection between each convolution layer in the encoder and the corresponding deconvolution layer with the same feature map size in the decoder.

Submask Extraction

A significant step of FRIH is to extract the submasks of the global mask. The number K of the submasks needs to be adaptive for different input images. Therefore, We apply CFSFDP clustering algorithm (Rodriguez and Laio 2014) to extract K submasks for each global mask. The original CFSFDP algorithm has two basic ideas, which are that cluster centers have a higher density than their neighbors, and are at a relatively large distance from points with higher densities (Rodriguez and Laio 2014). In our method, we cluster the pixels in the global mask by their corresponding pixel RGB values in the composite image. The local density ρ_i of pixel p_i is measured as:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{1}$$

where d_{ij} is the normalized euclidean distance between the (r, g, b) vectors of pixels p_i and p_j. d_c is the cutoff distance. We will introduce the setting of d_c in the experiments section. $\chi(x) = 1$ if $x < 0$ and $\chi(x) = 0$ otherwise. The minimum distance δ_i between pixel p_i and any other pixel p_j with higher density is defined as:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{2}$$

Note that if p_i has the highest density in the image, then $\delta_i = \max_j(d_{ij})$. In our method, according to the statistical observations on a large number of images, we find that

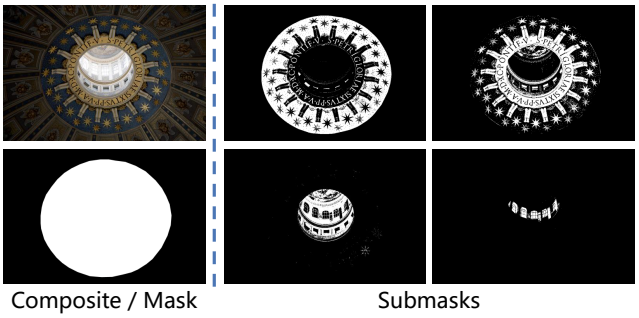

Figure 3: A representative case of the submask extraction. The origin mask (left) is divided into 4 submasks (right).

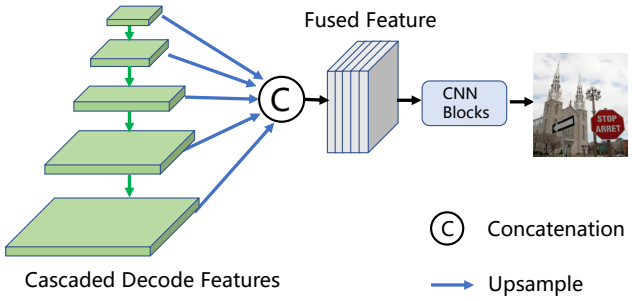

Figure 4: The fusion prediction module. Features from all the cascaded decoder layers are used to predict results.

the appropriate number of submasks for the foreground will not exceed 10 in the vast majority of cases. Therefore, we sort all pixels by δ. The 10 RGB values with the highest δ are considered to be the candidate cluster centers. Note that the RGB values of different candidate cluster centers should be different. For example, if all the pixels in the foreground have the same RGB value, there will be only 1 candidate cluster centers. The candidate cluster centers whose densities ρ do not exceed 10 will be considered as isolated outliers. Only the candidate cluster centers whose densities are higher than 10 are considered as the final cluster centers. Thus the cluster centers of each foreground mask are obtained adaptively. The number of the cluster centers K is in the range of 1 to 10.

When all the cluster centers are obtained, the remaining pixels are assigned to the same cluster as their nearest pixel of higher density. In this way, all the pixels in the foreground mask M_f are assigned into K clusters. For each composite image, we obtain K submasks $Subm^1$,$Subm^2$,...,$Subm^k$. Figure 3 displays a representative set of the generated submasks. In this case, $K = 4$.

Lightweight Cascaded Module

In this module, we concatenate the global coarsely adjusted image generated in the first stage and each submask respectively. All the concatenated pairs are fed into the lightweight cascaded module. We imitate the structure of U-Net (Ronneberger, Fischer, and Brox 2015) to construct the lightweight cascaded module. In the cascaded encoder, seven 4×4 convolution layers are used to extract different level features of the coarsely adjusted image. In the cascaded decoder part, different from the original decoder in U-Net, at each layer, we use a 1×1 convolution layer to fuse the features from three sources: the previous decoder layer, the corresponding encoder layer in the first stage (the yellow arrow in Figure 2), the corresponding encoder layer in the cascaded module. In this way, we not only further harmonize the image based on the results from first stage network, but also utilize the features from the original composite image. The details of this module are shown in Figure 2. Thus the output of the i-th cascaded decoder layer DC_i can be calculated by the following equation:

$$ET_i = Conv_{trans}(E_i) \tag{3}$$

$$F_i = Conv_{fuse}([DC_{i-1}, ET_i, EC_i]) \tag{4}$$

$$DC_i = Conv_{up}(F_i) \tag{5}$$

where E_i denotes the features from the first stage encoder. We use two 3×3 convolution layers as transfer function $Conv_{trans}$ to transfer E_i into ET_i to make it suitable for the cascaded module. EC_i denotes the output of the i-th layer in the cascaded encoder, which contains the information from the coarsely adjusted image. The operator [] means the concatenation operation and $Conv_{fuse}$ is a 1×1 convolution layer to fuse the features from three different sources. $Conv_{up}$ is a 4×4 transpose convolution layer to upsample and decode the fused feature F_i. In this way, we obtain the cascaded adjust features from the last layer of our cascaded module. Note that the channels of the features are much smaller than those in the original U-Net. The cascaded module together with the following embedded fusion prediction module has only 2.68 M parameters, only accounting for 22.4% of the parameters of the whole FRIH network. Therefore, our cascaded module is lightweight.

Fusion Prediction Module

In order to utilize the different levels of harmonization results comprehensively, we further design a fusion prediction module. In the previous image harmonization methods and traditional U-Net, they only use the features from the last decoder layer to predict the generated images, since the features from the former decoder layers have not been adjust to be close to the target enough. However, in our cascaded module, the input is the coarsely adjusted image. All the cascaded decoder layers have already fuse the information of the coarsely harmonized image. The outputs from different cascaded decoder layers are the features of images adjusted to different degrees. The similarity between the composite image and the ground-truth image varies among different sub-regions. Some sub-regions in the composite image are relatively similar to the ground-truth image, while others may be much far away from the target. Thus different sub-regions need different degrees of harmonization. As shown in Figure 4, we fuse features from all the cascaded decoder layers together to predict the harmonized result, which enables the prediction head to utilize features from different harmonization levels. Since the resolution of the features

from different cascaded decoder layers are different, we upsample all these feature maps to the resolution 256×256 and concatenate them together. Then we use two 3×3 and one 1×1 convolution layers to convert these fused feature maps to a 3-channel RGB image. As shown in Figure 2, we obtain K refined images from the input K pairs. The final output image is generated by combining these images according to the corresponding submasks.

Training Loss

Since we use base network to predict coarse results and a cascaded module to obtain refined images, for each image, the loss function L_{total} consists of two parts, L_{coarse} and L_{refine}. We use the following equation to calculate L_{total}:

$$L_{total} = L_{coarse} + L_{refine} \tag{6}$$

$$L_{coarse} = \sum_{h,w} \frac{||I_{h,w} - \hat{I}_{h,w}||_2^2}{max(Area_{mask}, A_{min})} \tag{7}$$

$$L_{refine} = \sum_{i=1}^{K} \sum_{h,w} \frac{||I_{h,w} - \hat{I}_{h,w}||_2^2 \cdot Subm_{h,w}^i}{max(Area_{Subm^i}, A_{min})} \tag{8}$$

where $\hat{I}$ is the prediction result. In equations 7 and 8, $\hat{I}$ represents the coarsely adjusted image in the first stage and the final output image in the second stage, respectively. It should be noted that in L_{refine}, we only focus on the submask area, so the loss is multiplied by the submask. Furthermore, We find that the images with small foreground masks are always hard examples and we want our model to learn more information from them. Therefore, we divide the loss by the area of the foreground mask. A_{min} is a constant, which is set to 100 in all the experiments. For those masks or submasks whose areas are smaller than A_{min}, we treat them as A_{min}.

Experiments

Datasets and Evaluation Metrics

To demonstrate the effectiveness of our FRIH, we conduct experiments on the public image harmonization dataset iHarmony4 (Cong et al. 2020). This dataset consists of four sub-datasets, including HCOCO, HAdobe5k, HFlickr and Hday2night. There are totally 65,742 training image pairs and 7,407 test image pairs in iHarmony4. All the image pairs are generated by modifying the specific foreground regions of the normal images, which are converted to corresponding inharmonious images in this way. We follow the same train-test split as DoveNet (Cong et al. 2020) in the experiments.

Following prior work (Tsai et al. 2017), we use Mean Squared Error (MSE) score and Peak Signal-to-Noise Ratio (PSNR) score on RGB channels to evaluate the image harmonization performance. To eliminate the influence of the foreground area size on the metric, we also introduced the foreground Mean Squared Error (fMSE) score (Cong et al. 2020). In addition, to make the evaluation criteria more aligned with human subjective standards, we also used the LPIPS metric (Zhang et al. 2018) for assessment. All the metrics are calculated based on the 256×256 resolution.

Implementation Details

We use Adam Optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to train our model for 180 epochs on 8 Tesla V100 GPUs. The initial learning rate is 0.008, which decays by 10 at epoch 160 and 175. The batchsize is 128. All the images are resized to 256×256 in both training and test process. We use horizontal flip and random size crop to augment the data during training. The whole FRIH network is trained end-to-end. The cutoff distance d_c is set to 0.1.

Comparison with the State-of-the-Art Methods

We compare our FRIH with other image harmonization methods on iHarmony4. Table 1 and Table 2 separately show the results in each sub-dataset and each foreground ratio range. From Table 1 and Table 2 we can find that:

(1) Our method achieves 38.19 PSNR and 23.98 MSE on iHarmony4 test set, which performs better than all the other image harmonization methods on iHarmony4 test set in a large margin. Compared to previous hamonization methods, the PSNR of FRIH has a certain improvement, proving that the fine-grained region-aware framework works well. It should be noted that iDIH (Sofiiuk, Popenova, and Konushin 2021) obtains higher performance than 37.08 dB when adding extra pre-trained semantic segmentation model, which all other methods do not use. It is unfair to make comparison so that we only show their results without the extra model.

(2) In Table 1, our method performs much better than existing methods on all the 4 sub-datasets, proving the robustness of our cascaded module and fusion prediction strategy.

(3) In Table 2, our method performs better than all the existing methods on the composite images with different foreground ratio, demonstrating that our submask extraction is well adaptive regardless of the foreground area size.

Sometimes the PSNR and MSE can not represent the impressions of humans. To prove FRIH can generate more harmonious images when evaluated by humans, we conduct a qualitative analysis and an user study in the supplementary.

Ablation Study

We compare the following models on iHarmony4 dataset to show the effectiveness of FRIH's parts:

- *Input composite*. We use the input composite image as the final output result, without harmonization operation.
- *Baseline*. We only use the base network (the first stage) to make a coarse-grained harmonization.
- *Baseline+Cascade Globalmask*. We add the lightweight cascaded module, but instead of the submasks, we feed the concatenation of the global mask and the coarsely adjusted image into the lightweight cascaded module.
- *Baseline+Cascade Submask*. We extract submasks adaptively for each foreground mask and feed the concatenation of each submask and the coarsely adjusted image into the lightweight cascaded module.
- *Baseline+Cascade Submask+Fusion (FRIH)*. This is the full version of our algorithm, including the submask extraction, lightweight cascaded module together with the

Sub-dataset	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
Metric	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑
Input composite	69.37	33.94	345.54	28.16	264.35	28.32	109.65	34.01	172.47	31.63
Lalonde (Lalonde and Efros 2007)	110.10	31.14	158.90	29.66	329.87	26.43	199.93	29.80	150.53	30.16
Xue (Xue et al. 2012)	77.04	33.32	274.15	28.79	249.54	28.32	190.51	31.24	155.87	31.40
Zhu (Zhu et al. 2015)	79.82	33.04	414.31	27.26	315.42	27.52	136.71	32.32	204.77	30.72
DIH (Tsai et al. 2017)	51.85	34.69	92.65	32.28	163.38	29.55	82.34	34.62	76.77	33.41
S^2AM (Cun and Pun 2020)	41.07	35.47	63.40	33.77	143.45	30.03	76.61	34.50	59.67	34.35
DoveNet (Cong et al. 2020)	36.72	35.83	52.32	34.34	133.14	30.21	54.05	35.18	52.36	34.75
ADFM (Hao et al. 2020)	-	36.87	-	34.99	-	33.36	-	34.31	-	35.86
BargainNet (Cong et al. 2021)	24.84	37.03	39.94	35.34	97.32	31.34	50.98	35.67	37.82	35.88
IIH (Guo et al. 2021b)	24.92	37.16	43.02	35.20	105.13	31.34	55.53	35.96	38.71	35.90
RainNet (Ling et al. 2021)	-	37.08	-	36.22	-	31.64	-	34.83	-	36.12
iDIH (Sofiiuk, Popenova, and Konushin 2021)	19.29	38.44	30.87	36.09	84.10	32.61	55.24	37.26	30.56	37.08
D-HT (Guo et al. 2021a)	16.89	38.76	38.53	36.88	74.51	33.13	53.01	37.10	30.30	37.55
S^2CRNET (Liang et al. 2022)	23.22	38.48	34.91	36.42	98.73	32.48	51.67	36.81	35.58	37.18
Harmonizer (Ke et al. 2022)	17.34	38.77	**21.89**	37.64	64.81	**33.63**	**33.14**	37.56	24.26	37.84
FRIH (ours)	**15.05**	**39.35**	23.61	37.69	68.93	33.48	42.78	**37.89**	**23.98**	**38.19**

Table 1: Comparison of image harmonization results in each sub-dataset on iHarmony4 test set.

Foreground ratios	0%-5%		5%-15%		15%-100%		All	
Metric	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
Input composite	28.51	1208.86	119.19	1323.23	577.58	1887.05	172.47	1387.30
Lalonde (Lalonde and Efros 2007)	41.52	1481.59	120.62	1309.79	444.65	1467.98	150.53	1433.21
Xue (Xue et al. 2012)	31.24	1325.96	132.12	1459.28	479.53	1555.69	155.87	1411.40
Zhu (Zhu et al. 2015)	33.30	1297.65	145.14	1577.70	682.69	2251.76	204.77	1580.17
DIH (Tsai et al. 2017)	18.92	799.17	64.23	752.86	228.86	768.89	76.77	773.18
S^2AM (Cun and Pun 2020)	15.09	623.11	48.33	540.54	177.62	592.83	59.67	594.67
DoveNet (Cong et al. 2020)	14.03	591.88	44.90	504.42	152.07	505.82	52.36	549.96
BargainNet (Cong et al. 2021)	10.55	450.33	32.13	359.49	109.23	353.84	37.82	405.23
RainNet (Ling et al. 2021)	11.66	550.38	32.05	378.69	117.41	389.80	40.29	469.60
iDIH (Sofiiuk, Popenova, and Konushin 2021)	8.38	366.32	25.39	287.02	89.44	297.94	30.56	330.45
FRIH (ours)	**6.89**	**305.28**	**19.88**	**226.45**	**70.05**	**205.83**	**23.98**	**252.63**

Table 2: Comparison of image harmonization results in each foreground ratio range on iHarmony4 test set.

fusion prediction strategy. We use features from all the cascaded decoder layers to predict the final results.

The ablation study results in each sub-dataset and each foreground ratio range on iHarmony4 are presented in Table 3 and Table 4 respectively, showing that:

(1) Both the lightweight cascaded module and fusion prediction strategy can increase PSNR and decrease MSE. *Baseline+Cascade Submask* performs significantly better than *Baseline* with the gain of 1.08 dB in PSNR. *Baseline+Cascade Submask+Fusion* outperforms *Baseline+Cascade Submask* with the gain of 0.29 dB in PSNR. It proves the effectiveness of the lightweight cascaded module together with the fusion prediction strategy. Compared to *Baseline*, our FRIH has a gain of 1.37 dB in PSNR and a decrease of 11.40 in MSE, demonstrating that our global-local framework is effective and all modules are reciprocal.

(2) *Baseline+Cascade Globalmask* only performs slightly better than *Baseline* with the gain of 0.16 dB in PSNR. While *Baseline+Cascade Submask* performs much better than *Baseline* with the gain of 1.08 dB in PSNR. It proves that our submask extraction strategy plays an key role in the lightweight cascasded module.

(3) Both the effectiveness of the lightweight cascaded module and the fusion prediction module is not obvious on Hday2night sub-dataset. By analysis, we find that the foreground areas in Hday2night are always sky, water surface or other similar categories with single and pure color. These areas are homogeneous. Therefore, it is hard to separate these foreground areas into different sub-regions, which reduces the effect of our cascaded module together with the lightweight cascaded module. From another perspective, it further proves the effectiveness of our method on foregrounds with different appearance patterns.

(4) The increase of PSNR between *Baseline* and *Baseline+Cascade Submask* on HAobe5k sub-dataset achieves 1.63, which is the largest among 4 sub-datasets in Table 3. We think the reason is that the foreground masks in HAdobe5k are much bigger than other three sub-datasets, which means these bigger foreground masks are more likely to contain disparate sub-regions or different appearance pat-

Sub-dataset	HCOCO		HAdobe5k		HFlickr		Hday2night		All	
Metric	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑	MSE↓	PSNR↑
Input composite	69.37	33.94	345.54	28.16	264.35	28.32	109.65	34.01	172.47	31.63
Baseline	21.38	38.24	37.15	35.79	90.12	32.25	50.12	37.45	35.38	36.82
Baseline+Cascade Globalmask	20.05	38.39	35.52	35.98	87.29	32.40	48.91	37.53	33.82	36.98
Baseline+Cascade Submask	16.52	39.09	26.21	37.42	75.82	33.05	44.21	37.78	26.31	37.90
Baseline+Cascade Submask+Fusion	**15.05**	**39.35**	**23.61**	**37.69**	**68.93**	**33.48**	**42.78**	**37.89**	**23.98**	**38.19**

Table 3: Ablation study in each sub-dataset on iHarmony4 test set.

Foreground ratios	0%-5%		5%-15%		15%-100%		All	
Metric	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓	MSE↓	fMSE↓
Input composite	28.51	1208.86	119.19	1323.23	577.58	1887.05	172.47	1387.30
Baseline	10.21	421.53	29.68	336.32	109.21	345.19	35.38	380.12
Baseline+Cascade Globalmask	9.89	401.18	27.85	319.47	99.35	320.71	33.82	366.23
Baseline+Cascade Submask	7.28	322.86	21.42	244.15	77.82	230.47	26.31	268.52
Baseline+Cascade Submask+Fusion	**6.89**	**305.28**	**19.88**	**226.45**	**70.05**	**205.83**	**23.98**	**252.63**

Table 4: Ablation study in each foreground ratio range on iHarmony4 test set.

d_c	0.01	0.05	0.1	0.2	0.3	0.4
MSE↓	25.26	24.49	**23.98**	24.76	25.41	25.90
PSNR↑	38.09	38.17	**38.19**	38.15	38.08	38.02

Table 5: Comparative experiment for d_c on iHarmony4.

	DoveNet	IIH	RainNet	FRIH (ours)
PSNR↑	34.75	35.90	36.12	**38.19**
Param (M)↓	54.76	40.83	54.75	**11.98**

Table 6: Comparison of performance and efficiency between our FRIH and other methods on iHarmony4.

terns. The baseline model treats them in the same way while the cascaded module can adjust them adaptively, leading to a significantly increase of PSNR and decrease of MSE.

(5) The increase between *Baseline* and *Baseline+Cascade Submask* of images whose foreground mask sizes are larger than 15% is also the biggest in Table 4. The reason is the same as why the increase in the sub-dataset HAdobe5k is the biggest. Both (4) and (5) prove that our submasks-based cascaded module is efficient for image harmonization task, especially for those images with large foreground masks.

Moreover, the cutoff distance d_c in the submask extraction module is an important hyper-parameter in FRIH. It determines the granularity of the submasks division. We conduct comparative experiment for d_c on iHarmony4. As shown in Table 5. d_c is set to 0.01, 0.05, 0.1, 0.2, 0.3 and 0.4, respectively. When $d_c = 0.1$, our FRIH achieves the best performance. Thus we set $d_c = 0.1$ in all the other experiments in this paper. When $d_c = 0.01$, the performance is lower than $d_c = 0.1$, because the clustering is so fine-grained that many noisy isolated outliers are generated and have negative effect on the selection of cluster centers. When $d_c = 0.3$ and $d_c = 0.4$, the performance also drops compared with $d_c = 0.1$, which is caused by the too coarse-grained clustering. Several color blocks, which are not similar, are clustered into a same cluster. However, The PSNR in all these experiments exceeds 38, proving the robustness of our method.

Comparison of Model Size and Computation

We also compare the model size and computation of our FRIH with existing methods. In Table 6, compared to other methods, our FRIH achieves much higher performance in PSNR. However, the model size of our FRIH (11.98 M) is lower than other methods. In these experiments, our FRIH achieves the best performance with the smallest model size, demonstrating the effectiveness and efficiency.

Conclusion

We propose a two-stage fine-grained region-aware image harmonization framework, which is simple, novel and effective. In the first stage, the whole input foreground mask is used to make a global coarse-grained harmonization. In the second stage, we adaptively cluster the input foreground mask into several submasks by the corresponding pixel RGB values. Each submask and the coarsely adjusted image are concatenated respectively and fed into the lightweight cascaded module. Moreover, we further design a fusion prediction module by fusing features from all the cascaded decoder layers to generate the final result. It addresses the problem that existing methods ignore the difference of each color block and lose some specific details. Extensive experiments demonstrate the effectiveness and efficiency of our FRIH and its superiority over the state-of-the-art competitors.

References

Azadi, S.; Pathak, D.; Ebrahimi, S.; and Darrell, T. 2020. Compositional gan: Learning image-conditional binary composition. *International Journal of Computer Vision*, 128(10): 2570–2585.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Chen, B.-C.; and Kae, A. 2019. Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8415–8424.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4983.

Cheng, Y.; Gan, Z.; Li, Y.; Liu, J.; and Gao, J. 2020. Sequential attention GAN for interactive image editing. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4383–4391.

Cohen-Or, D.; Sorkine, O.; Gal, R.; Leyvand, T.; and Xu, Y.-Q. 2006. Color harmonization. In *Proceedings of SIGGRAPH*.

Cong, W.; Niu, L.; Zhang, J.; Liang, J.; and Zhang, L. 2021. Bargainnet: Background-Guided Domain Translation for Image Harmonization. In *IEEE International Conference on Multimedia and Expo*.

Cong, W.; Zhang, J.; Niu, L.; Liu, L.; Ling, Z.; Li, W.; and Zhang, L. 2020. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Cun, X.; and Pun, C.-M. 2020. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Guo, Z.; Chen, Z.; Yu, T.; Chen, J.; and Liu, S. 2019. Progressive image inpainting with full-resolution residual network. In *Proceedings of the 27th acm international conference on multimedia*, 2496–2504.

Guo, Z.; Guo, D.; Zheng, H.; Gu, Z.; Zheng, B.; and Dong, J. 2021a. Image Harmonization With Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14870–14879.

Guo, Z.; Zheng, H.; Jiang, Y.; Gu, Z.; and Zheng, B. 2021b. Intrinsic Image Harmonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hao, G.; et al. 2020. Image Harmonization with Attention-based Deep Feature Modulation. In *The British Machine Vision Conference*.

Jia, J.; Sun, J.; Tang, C.-K.; and Shum, H.-Y. 2006. Drag-and-drop pasting. *ACM Transactions on graphics*.

Jiang, Y.; Zhang, H.; Zhang, J.; Wang, Y.; Lin, Z.; Sunkavalli, K.; Chen, S.; Amirghodsi, S.; Kong, S.; and Wang, Z. 2021. SSH: A Self-Supervised Framework for Image Harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4832–4841.

Ke, Z.; Sun, C.; Zhu, L.; Xu, K.; and Lau, R. W. 2022. Harmonizer: Learning to perform white-box image and video harmonization. In *European Conference on Computer Vision*, 690–706. Springer.

Lalonde, J.-F.; and Efros, A. A. 2007. Using color compatibility for assessing image realism. In *IEEE/CVF International Conference on Computer Vision*.

Liang, J.; Cun, X.; Pun, C.-M.; and Wang, J. 2022. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. In *European Conference on Computer Vision*, 334–349. Springer.

Ling, J.; Xue, H.; Song, L.; Xie, R.; and Gu, X. 2021. Region-aware Adaptive Instance Normalization for Image Harmonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Liu, D.; Puri, R.; Kamath, N.; and Bhattacharya, S. 2020. Composition-aware image aesthetics assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 3569–3578.

Liu, H.; Shen, X.; Shang, F.; Ge, F.; and Wang, F. 2019. CU-Net: Cascaded U-Net with loss weighted sampling for brain tumor segmentation. In *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*, 102–111. Springer.

Pérez, P.; Gangnet, M.; and Blake, A. 2003. Poisson image editing. In *Proceedings of SIGGRAPH*.

Pitié, F.; and Kokaram, A. 2007. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *European Conference on Computer Vision*.

Pitie, F.; Kokaram, A. C.; and Dahyot, R. 2005. N-dimensional probability density function transfer and its application to color transfer. In *IEEE/CVF International Conference on Computer Vision*.

Qiu, H.; Xiao, C.; Yang, L.; Yan, X.; Lee, H.; and Li, B. 2020. SemanticAdv: Generating Adversarial Examples via Attribute-conditioned Image Editing. In *European Conference on Computer Vision*, 19–37. Springer.

Rodriguez, A.; and Laio, A. 2014. Clustering by fast search and find of density peaks. *science*, 344(6191): 1492–1496.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention*.

Sofiiuk, K.; Popenova, P.; and Konushin, A. 2021. Foreground-aware Semantic Representations for Image Harmonization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1620–1629.

Song, S.; Zhong, F.; Qin, X.; and Tu, C. 2020. Illumination harmonization with gray mean scale. In *Computer Graphics International Conference*, 193–205. Springer.

Sunkavalli, K.; Johnson, M. K.; Matusik, W.; and Pfister, H. 2010. Multi-scale image harmonization. *ACM Transactions on Graphics*.

Tao, M. W.; Johnson, M. K.; and Paris, S. 2013. Error-tolerant image compositing. *International journal of computer vision*.

Tsai, Y.-H.; Shen, X.; Lin, Z.; Sunkavalli, K.; Lu, X.; and Yang, M.-H. 2017. Deep image harmonization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, J.; Wang, C.; Huang, Q.; Shi, Y.; Cai, J.-F.; Zhu, Q.; and Yin, B. 2020. Image inpainting based on multi-frequency probabilistic inference model. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1–9.

Wu, R.; Zhang, G.; Lu, S.; and Chen, T. 2020. Cascade efgan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5021–5030.

Xue, S.; Agarwala, A.; Dorsey, J.; and Rushmeier, H. 2012. Understanding and improving the realism of image composites. *ACM Transactions on graphics*.

Zhan, F.; Lu, S.; Zhang, C.; Ma, F.; and Xie, X. 2020. Adversarial image composition with auxiliary illumination. In *Proceedings of the Asian Conference on Computer Vision*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.-Y.; Krahenbuhl, P.; Shechtman, E.; and Efros, A. A. 2015. Learning a discriminative model for the perception of realism in composite images. In *IEEE/CVF International Conference on Computer Vision*.