

Harmonizer: Learning to Perform White-Box Image and Video Harmonization

Zhanghan Ke¹, Chunyi Sun², Lei Zhu¹, Ke Xu¹, and Rynson W.H. Lau^{1†}

¹ Department of Computer Science, City University of Hong Kong

² Australian National University

Abstract. Recent works on image harmonization solve the problem as a pixel-wise image translation task via large autoencoders. They have unsatisfactory performances and slow inference speeds when dealing with high-resolution images. In this work, we observe that adjusting the input arguments of basic image filters, *e.g.*, brightness and contrast, is sufficient for humans to produce realistic images from the composite ones. Hence, we frame image harmonization as an image-level regression problem to learn the arguments of the filters that humans use for the task. We present a *Harmonizer* framework for image harmonization. Unlike prior methods that are based on black-box autoencoders, Harmonizer contains a neural network for filter argument prediction and several white-box filters (based on the predicted arguments) for image harmonization. We also introduce a cascade regressor and a dynamic loss strategy for Harmonizer to learn filter arguments more stably and precisely. Since our network only outputs image-level arguments and the filters we used are efficient, Harmonizer is much lighter and faster than existing methods. Comprehensive experiments demonstrate that Harmonizer surpasses existing methods notably, especially with high-resolution inputs. Finally, we apply Harmonizer to video harmonization, which achieves consistent results across frames and 56 *fps* at 1080P resolution.

1 Introduction

Extracting the foreground from one image and compositing it onto a background image is a popular operation in vision applications, *e.g.*, image editing [29,2] and stitching [44,45]. In order for the composite image to be more realistic, *i.e.*, cannot be easily distinguished by humans, the image harmonization task is introduced to remove the inconsistent appearances between the foreground and background. This task is challenging because many conditions, such as lighting and imaging device being used, can affect object visual appearances [46,3], and humans are sensitive to even fine inharmony in appearances [23,42].

Traditional methods [47,31,7,18,23,36,28] focus on matching the hand-crafted statistics between foreground and background regions, disregarding the semantic information which is vital for eliminating the large appearance gap. Recent

[†] Rynson W.H. Lau is the corresponding author.

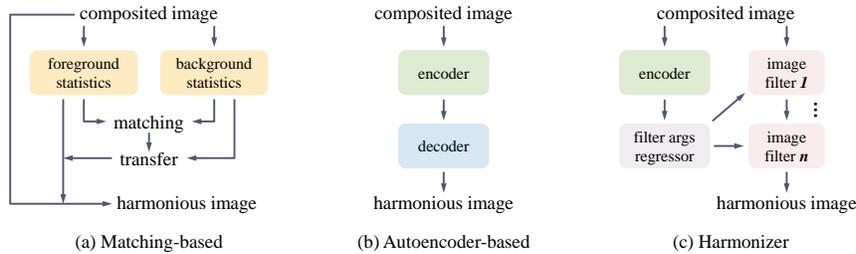


Fig. 1. Comparison of Harmonization Frameworks. (a) Traditional matching-based methods transfer the background appearance to foreground regions based on hand-crafted statistics. (b) Autoencoder-based methods use black-box models to obtain harmonious images via pixel-wise image translation. (c) Our Harmonizer regresses image-level filter arguments to perform image harmonization in a white-box manner.

deep-learning based methods [39,9,14,16,35,19,12] leverage the strong semantic representation capability of autoencoders [34] to lower the appearance gap. They regard image harmonization as a pixel-wise image translation task [17,40] from a composite image to a harmonious version. Although they have achieved notable success, they also suffer from three key problems. First, their performances are unsatisfactory at high resolutions. Since using high-resolution images for training requires a huge amount of resources, these methods usually train and evaluate at low resolution. Second, these methods are not suitable for mobile devices or real-time applications, due to their high computational overheads. The size of recent autoencoder-based models [9,27,8,13] is larger than 100 MB, and their inference speed at 1080P (Full HD) resolution is only ~ 10 *fps* on a RTX3090 GPU. Third, the images generated by these methods may not be consistent with the input images in terms of textures/details, *i.e.*, the original image contents may be changed, because neural networks are still essentially black-box models.

To design an efficient strategy for resolution-independent image harmonization in a white-box manner, we conduct a user study to explore how humans perform this task. We observe that humans are able to produce realistic images by adjusting the input arguments of some basic image filters, such as brightness and contrast. These filters also do not suffer from the three aforementioned problems, *i.e.*, resolution-dependence, inefficiency, and black-box manner. Motivated by our observation, we formulate the image harmonization task as an image-level regression problem to learn the arguments of the filters used by humans, and present a *Harmonizer* framework for the task. The key idea of our design is to combine a neural network and white-box filters for image harmonization, rather than just using black-box autoencoders. Specifically, in Harmonizer, the network contains a backbone encoder and a regressor for filter argument prediction, while the white-box filters use the predicted arguments to harmonize the input composite images. Fig. 1 summarizes the main differences between Harmonizer and existing frameworks.

To learn filter arguments more stably and precisely, we need to further consider two problems. First, the filter arguments are not easy to optimize simulta-

neously since they may affect each other. For example, if we adjust the brightness first before adjusting the highlight, we should consider the brightness argument when regressing the highlight argument. We note that utilizing a straightforward multiple-head regressor to predict each filter argument independently has unsatisfactory performances. We solve this problem by introducing a cascade regressor to predict each filter argument based on the features of the preceding filter arguments. Second, the loss of each filter output would accumulate all errors from the preceding filters, causing the regressor to bias towards some filters. We address this problem by introducing a dynamic loss strategy, which can balance the losses and helps Harmonizer pay more attention to the filters that are more difficult to learn. Besides, we design a simple but effective exponential moving average (EMA) based strategy to adapt Harmonizer to video harmonization.

We conduct extensive experiments to evaluate Harmonizer. The results on the iHarmony4 benchmark [9] demonstrate that Harmonizer outperforms prior state-of-the-art by a large margin. Harmonizer also has clear advantages in terms of model size and inference speed. Our ablation study verifies the effectiveness of each component of Harmonizer. For video harmonization, Harmonizer obtains consistent results across frames and an inference speed of 56 *fps* at 1080P resolution on a RTX3090 GPU.

2 Related Works

2.1 Image Harmonization

For an image composited of foreground image F with foreground mask M and background image B , the image harmonization task optimizes a harmonization function \mathcal{H} that processes the foreground region MF in order to match with the visual appearance of B , *i.e.*, creating a natural image I , as:

$$I = \mathcal{H}(MF) + (1 - M)B. \quad (1)$$

Most traditional algorithms proposed \mathcal{H} functions that concentrated on matching low-level appearance statistics, including color distributions [31,33,30,42], color templates [7], and gradient domain [18,29,38]. Some works further combined multi-scale statistics [36] or considered the visual realism of images [23,20].

In recent years, many methods based on CNNs have been proposed with notable successes. These works regarded image harmonization as a pixel-wise image translation task, and their \mathcal{H} functions are implemented based on autoencoders. For example, Tsai *et al.* [39] trained an end-to-end autoencoder to explore high-level semantics. Cun *et al.* [10] introduced a spatial-separated attention module to leverage low-level appearances. Cong *et al.* [5,4] focused on finding more effective methods to guide the processing of the foreground using the information from the background. Ling *et al.* [27] related image harmonization with background-to-foreground style transfer. Guo *et al.* [13] considered the intrinsic image characteristics to handle reflectance and illumination. Guo *et*

al. [12] replaced the CNN encoder with a Transformer to capture global background context.

In spite of the success, all the aforementioned methods suffer from poor performances and slow inference speeds at high resolution, due to the low-resolution images used in the training process and the high computational overheads of autoencoders. Instead, in this work, we formulate the image harmonization task as an image-level regression problem, and our proposed Harmonizer can solve the task with a consistent inference speed at high resolutions with negligible performance degradation.

2.2 White-box Image Editing

Recently, some works combined neural networks with human understandable (*i.e.*, white-box) algorithms for image editing. These methods usually have more stable performance than using only black-box neural networks. In addition, while the results from black-box neural networks may not be invertible, white-box algorithms allow users to further edit the images or undo any unwanted operations. For example, Yan *et al.* [43] used model predictions to adjust pixel values. Zou *et al.* [48] proposed a generative framework with a renderer/blender to simulate the human painting process. Hu *et al.* [15] applied differentiable image operators for photo post-processing based on reinforcement learning. Wang *et al.* [41] finished cartoon stylization by tuning the representations decomposed from the images.

In the image harmonization task, existing deep learning based methods are all based on black-box autoencoders [39,9,8,27,13,12], except for a concurrent work that attempts to support high-resolution inputs [25]. In contrast, our proposed Harmonizer combines a neural network with image filters to perform image harmonization in a white-box manner.

The work most relevant to ours is probably Hu *et al.* [15]. However, they used reinforcement learning to predict both types and arguments of filters. Besides, their method regresses only one filter in each step and may perform the same filter multiple times, resulting in a slow inference speed. Instead, we regress the arguments of a set of filters simultaneously and performs each filter only once, avoiding redundant filter operations and greatly improving efficiency.

3 Harmonizer

3.1 Design Motivation

Harmonizer aims to address image harmonization from a new perspective - combining neural networks with a white-box strategy. Since the white-box strategy that we select should be understood by humans, we first conduct a two-stage user study to analyze how humans perform image harmonization.

In the first stage, we investigate the white-box strategy humans use for image harmonization. We ask 5 experts who work in the image editing field (2 photographers, 2 designer, and 1 painter) to process composite images with Photoshop.

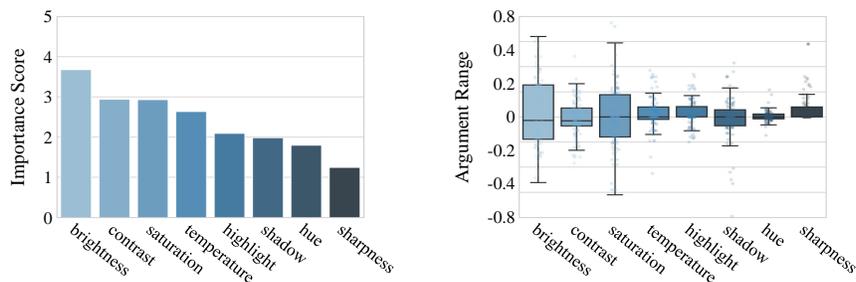


Fig. 2. Statistics of Our User Study. *Left:* We sort image filters by their average importance scores. *Right:* For each filter, we use a Gaussian to fit its input arguments from the participants and visualize the argument distribution via Boxplots.

We note that they accomplish this task mainly by modifying some image properties through tools that can be split into a set of image filters. For example, the “Levels” tool in Photoshop combines the highlight filter, the shadow filter, and the contrast filter. So, the 5 experts are essentially using image filters for image harmonization. In Harmonizer, we select the white-box strategy the same as the 5 experts: adjusting the arguments of appropriate filters to edit the foreground to match the background.

In the second stage, we study the importance of different filters in humans’ perception, and the value ranges that humans tune the filter arguments. We build an image harmonization system based on the filters used in the first stage. We invite 26 participants. For each of them, our system will display 10 composite images, including 5 images that are identical among all participants and 5 images randomly selected for each participant. For each composite image, the participants are required to adjust the given filters to make it looks natural. Meanwhile, they should give an importance score for each filter, indicating its role in processing the composite images. The score values are between 1 and 5. The higher the score, the more important the filter is. We record the importance scores and the filter arguments input by the participants for statistics. As shown in Fig. 2, the average importance scores of filters (Fig. 2 *Left*) guide us to choose the filters with high scores, *i.e.*, the filters that are more important in humans’ perception, for Harmonizer. The distributions of filter arguments (Fig. 2 *Right*) guide us to set appropriate value ranges for filter arguments, *i.e.*, the value ranges that humans use.

Based on the user study above, we determine the white-box strategy used in Harmonizer (the first stage). We also understand which filters are important and the appropriate argument ranges for the filters (the second stage).

3.2 Architecture

As shown in Fig. 3, the framework of Harmonizer contains a backbone encoder \mathcal{E} , a regressor \mathcal{R} , and a set of image filters $\mathcal{F} = \{\mathcal{F}_1, \dots, \mathcal{F}_k\}$, where k indicates the number of filters. The backbone \mathcal{E} in Harmonizer is EfficientNet-B0 [37]. Given

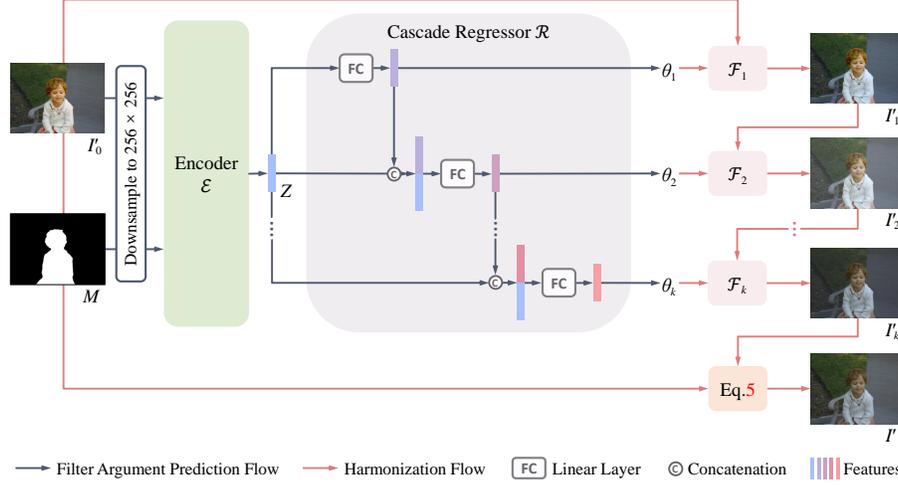


Fig. 3. The Harmonizer Framework. For an input image I'_0 with its foreground mask M , Harmonizer uses a neural network ($\mathcal{E}+\mathcal{R}$) to regress k image filter arguments $\theta = \{\theta_1, \dots, \theta_k\}$ (*i.e.*, the Filter Argument Prediction Flow). The image filters $\mathcal{F} = \{\mathcal{F}_1, \dots, \mathcal{F}_k\}$ in Harmonizer are then executed in sequence with the predicted arguments θ to obtain the output image I' (*i.e.*, the Harmonization Flow).

a composite image I'_0 and its corresponding foreground mask M , Harmonizer first downsamples them to the resolution of 256×256 and inputs them to \mathcal{E} to extract image features Z (with 160 channels), as:

$$Z = \mathcal{E}(I'_0, M). \quad (2)$$

Then, Harmonizer processes Z by global pooling and uses \mathcal{R} to regress filter arguments θ from it, as:

$$\theta = \{\theta_1, \dots, \theta_k\} = \mathcal{R}(Z), \quad \text{where } \theta_i \in [-1, 1], \quad i = 1, \dots, k. \quad (3)$$

With θ , Harmonizer executes the k filters in sequence on I'_0 , as:

$$I'_i = \mathcal{F}_i(I'_{i-1}, \theta_i), \quad i = 1, \dots, k. \quad (4)$$

Finally, the harmonious image I' is created by:

$$I' = MI'_k + (1 - M)I'_0. \quad (5)$$

Eq.5 ensures that the background regions in I' are the same as I'_0 , *i.e.*, the background pixels are not changed.

To balance the performance and the speed, we have also identified the preferred number of filters k and which k filters to use in Harmonizer. Our evaluations show that setting $k = 6$ is able to satisfy the real-time requirement (Table 4 *Right*). The six most important filters that we have selected based on human

perception (Fig. 2 *Left*) for Harmonizer include brightness, contrast, saturation, color temperature, highlight, and shadow.

Cascade Regressor. Predicting the arguments for k filters can be considered as a multi-task problem. One straightforward solution is to obtain each filter argument θ_i independently through a fully connected regressor \mathcal{R} with k heads $\{\mathcal{R}_1, \dots, \mathcal{R}_k\}$, as:

$$\theta_i = \mathcal{R}_i(Z), \quad i = 1, \dots, k. \quad (6)$$

However, Eq. 6 does not take into account the relationship between the filters. For example, both the brightness filter (with θ_b) and the highlight filter (with θ_h) will process the pixels with large pixel values. If we independently predict θ_b and θ_h from Z and constrain them with ground truth, both of them will attempt to make the composite input image look harmonious. As a result, the effects of these two filters will be accumulated in the output image, leading to an unsatisfactory result. To address this problem, we introduce a cascade regressor that uses the feature vector of the preceding filter arguments as conditions when regressing an argument θ_i , as:

$$\begin{aligned} \theta_1 &= \mathcal{R}_1(Z), \\ \theta_i &= \mathcal{R}_i(Z | \theta_{i-1}) = \mathcal{R}_i(Z | \theta_{i-1}, \dots, \theta_1), \quad i = 2, \dots, k. \end{aligned} \quad (7)$$

In practice, after predicting a filter argument, we concatenate its feature vector with Z to regress the next argument.

3.3 Training Strategy

We generate the composite input images from natural images for training. Since Harmonizer executes the k filters in a specific order $\mathcal{F}_1 \rightarrow, \dots, \rightarrow \mathcal{F}_k$ on the composite image I'_0 , we reverse this filter order to $\mathcal{F}_k \rightarrow, \dots, \rightarrow \mathcal{F}_1$ to create I'_0 from a natural image I , as:

$$\begin{aligned} I_k &= I, \\ I_{k-i} &= F_{k-i+1}(I_{k-i+1}, \xi_{k-i+1}), \quad i = 1, \dots, k, \\ I'_0 &= I_0, \end{aligned} \quad (8)$$

where ξ_{k-i+1} is the input arguments inside the range of $[-1, 1]$. However, some filters (*e.g.*, the color temperature filter) are sensitive to the input arguments and may drastically change the image appearance with even a small change in argument value, resulting in an irreversible I'_0 , *i.e.*, we may not be able to recover I . To alleviate this problem, we propose to reduce the range of the argument values when generating I'_0 . As shown in Fig. 2 *Right*, our user study provides a rough argument range for each filter, which can guarantee the reversibility of the composite images in most cases. Therefore, we sample the input argument ξ_i for the filter \mathcal{F}_i from a Gaussian distribution \mathcal{G}_i , as:

$$\xi_i = \mathcal{G}_i(m_i, v_i), \quad i = 1, \dots, k, \quad (9)$$

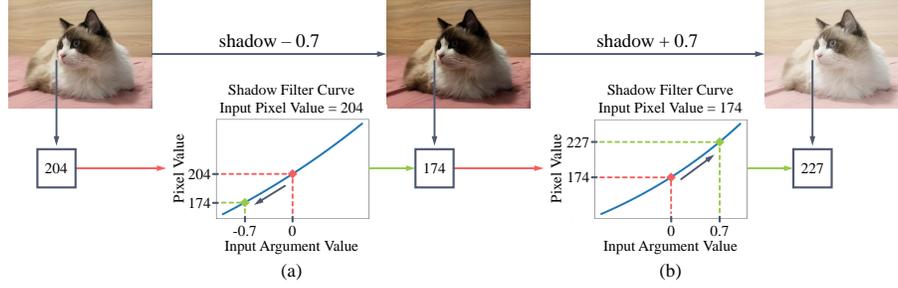


Fig. 4. Asymmetric Filter Operations. For the shadow filter that we use, if we (a) adjust the shadow with an argument of -0.7, the pixel value of 204 will drop to 174. After that, if we (b) adjust the shadow with an argument of 0.7, the pixel value of 174 will increase to 227, which is not equal to the original pixel value of 204.

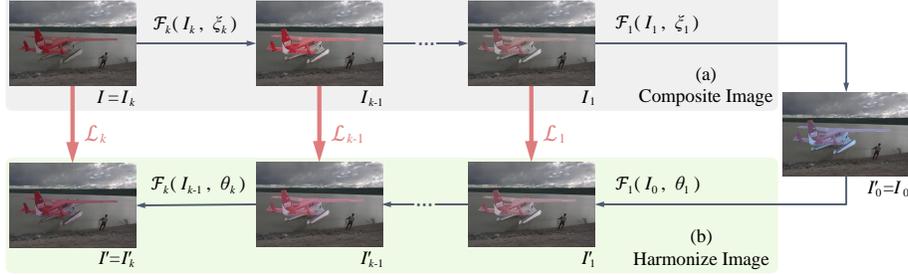


Fig. 5. Optimizing Filter Arguments θ . For (a) a composite image I'_0 generated from a natural image I using the filter arguments $\xi = \{\xi_1, \dots, \xi_k\}$, Harmonizer tends to (b) predict a set of arguments θ to recover a harmonious image I' from I'_0 . For some filters, the ground truth of θ_i is unknown. Hence, we optimize θ through the loss \mathcal{L}_i between each filter output I'_i and its corresponding composite image I_i .

where m_i and v_i are the mean and variance from Fig. 2 *Right*, respectively.

Note that some of the filters used are non-linear, and their operations are asymmetric. Here, we take the shadow filter \mathcal{F}_s as an example. As illustrated in Fig. 4, if we use the shadow filter \mathcal{F}_s with an input argument ξ_s to adjust an image I , we may not recover I using \mathcal{F}_s with argument $-\xi_s$. Therefore:

$$I \neq \mathcal{F}_s(\mathcal{F}_s(I, \xi_s), -\xi_s). \quad (10)$$

Hence, $-\xi_s$ cannot be used as the ground truth of the shadow filter argument θ_s predicted by Harmonizer. As shown in Fig. 5, instead of regressing θ directly, we optimize each θ_i through the L2 loss between the filter output I'_i and its corresponding composite image I_i (calculated in Eq. 8), as:

$$\mathcal{L}_i = M \|I'_i - I_i\|_2 = M \|\mathcal{F}_i(I'_{i-1}, \theta_i) - I_i\|_2, \quad i = 1, \dots, k. \quad (11)$$

Here the foreground mask M constrains the loss only on the foreground regions. We apply a loss on each output to ensure gradient propagation for the filters

in the front. For the composite input image generated by GAN [11] (following [9,8,12] *etc.*) rather than Eq. 8, we only apply the loss \mathcal{L}_k on the final output.

Dynamic Loss Strategy. During training, the Loss \mathcal{L}_i usually increases with the filter index i due to the inaccurate I_{i-1} from the preceding filters, which may bias the regressor towards later filters. We introduce a dynamic strategy to balance the losses. We first subtract the errors accumulated by the preceding filters from \mathcal{L}_i . We then normalize \mathcal{L}_i to enhance the loss of filters that introduce larger errors. Formally, the loss \mathcal{L}_i is dynamically reweighted by:

$$\tilde{\mathcal{L}}_i = \max\left(\frac{\mathcal{L}_i - \mathcal{L}_{i-1}}{\mathcal{L}_k}, 0\right), \quad i = 1, \dots, k, \quad (12)$$

Note that we detach the gradients at the denominator \mathcal{L}_k . If $\mathcal{L}_i < \mathcal{L}_{i-1}$, we reweight \mathcal{L}_i to 0 to focus on optimizing \mathcal{L}_{i-1} as we consider \mathcal{F}_i work well for the current input. The final training loss for Harmonizer is:

$$\mathcal{L} = \mu \sum_{i=1}^k \tilde{\mathcal{L}}_i. \quad (13)$$

where μ is used to rescale \mathcal{L}_i to ensure sufficient gradients for backpropagation.

3.4 Video Harmonizer

Applying existing image harmonization algorithms individually on each video frame often leads to flickering of the foreground in the output sequence. Although some video processing methods [4,5,22,24] have been proposed to encourage the prediction consistency across video frames, they require a long processing time or additional modules for training. Therefore, obtaining stable results in real-time video harmonization is an unexplored problem.

We introduce here a simple but effective strategy for adapting Harmonizer to video harmonization. The idea behind our strategy is to ensure that the predicted filter arguments change smoothly across frames. We achieve this by smoothing the predicted arguments θ with exponential moving average (EMA), as:

$$\bar{\theta}^t = (1 - \alpha) \bar{\theta}^{t-1} + \alpha \theta^t, \quad (14)$$

where t is the frame index, and $\alpha = 0.9$ is an EMA coefficient.

4 Experiments

In this section, we first introduce the datasets, metrics, and training details for our experiments. We then compare Harmonizer with existing image harmonization methods (Sec. 4.1). We also show the effectiveness of adapting Harmonizer to video harmonization (Sec. 4.2). We further conduct ablation experiments to evaluate the effectiveness of individual components in Harmonizer (Sec. 4.3). Finally, we demonstrate the advantages of Harmonizer in real-world image/video harmonization applications through user studies (Sec. 4.4).

Table 1. Quantitative Comparison on iHarmony4 at 256×256 Resolution. All metrics are computed following the previous works. \uparrow indicates the higher the better, while \downarrow indicates the lower the better.

Dataset	Metric	DIH[39]	S2AM[10]	DOVE[9]	BARG[8]	IntrIH[13]	IHT[12]	Our
HAdobe5k	MSE \downarrow	92.65	63.40	52.32	39.94	43.02	47.96	21.89
	fMSE \downarrow	593.03	404.62	380.39	359.49	284.21	321.14	170.05
	PSNR \uparrow	32.28	33.77	34.34	35.34	35.20	36.10	37.64
HFlickr	MSE \downarrow	163.38	143.45	145.21	97.32	105.13	88.41	64.81
	fMSE \downarrow	1099.13	785.65	985.79	769.02	716.60	617.26	434.06
	PSNR \uparrow	29.55	30.03	29.75	31.34	31.34	32.37	33.63
HCOCO	MSE \downarrow	51.85	41.07	36.72	24.84	24.92	20.99	17.34
	fMSE \downarrow	798.99	542.06	551.01	489.94	416.38	377.11	298.42
	PSNR \uparrow	34.69	35.47	35.83	37.03	37.16	37.87	38.77
Hday2night	MSE \downarrow	82.34	76.61	56.92	50.98	55.53	58.14	33.14
	fMSE \downarrow	1129.40	989.07	1067.19	853.61	797.04	823.68	542.07
	PSNR \uparrow	34.62	34.50	35.53	35.88	35.96	36.38	37.56
All	MSE \downarrow	76.77	59.67	52.36	37.82	38.71	37.07	24.26
	fMSE \downarrow	778.41	537.23	541.53	513.16	400.29	395.66	280.51
	PSNR \uparrow	33.41	34.35	34.75	35.88	35.90	36.71	37.84

Datasets. Following the recent works, we conduct our experiments on the iHarmony4 benchmark [9], which contains four subsets: HCOCO, HAdobe5k, HFlickr, and Hday2night. Each sample in iHarmony4 consists of a natural image, a foreground mask, and a composite image (with the foreground generated by GAN [11]). During training, we also create the composite images via Eq. 8. Note that this is a data augmentation method, without using any extra data.

Metrics. We evaluate the image harmonization performance by Mean Square Error (MSE), foreground MSE (fMSE), and Peak Signal-to-Noise Ratio (PSNR). fMSE calculates MSE only on the foreground regions rather than the whole image, as image harmonization does not change the background appearance.

Training. We train Harmonizer by Adam for 60 epochs. With a batch size of 16, the initial learning rate is set to $3e^{-4}$ and is multiplied by 0.1 after every 25 epochs. We set μ (in Eq. 12) to 10. In all experiments, except the ablation on the number of filters, we use Harmonizer with the 6 filters stated in Sec. 3.2.

4.1 Comparison with State-of-the-arts

We compare Harmonizer with recently proposed methods, including DIH [39], S2AM [10], DOVE [9], BARG [8], IntrIH [13], and IHT [12]. We use the pre-trained models released by their authors for evaluation. We first follow the prior works to evaluate all methods at low resolution, *i.e.*, the output harmonious images and the ground truths will be resized to 256×256 for metric calculation. As shown in Table 1, Harmonizer outperforms the existing methods on all four subsets of iHarmony4. Notably, compared to the state-of-the-art method, Harmonizer reduces the average MSE across all subsets by 35%.

For practical applications, which typically use higher image resolutions, the quantitative results at 256×256 resolution as shown above may not reflect

Table 2. Quantitative Comparison on iHarmony4 at High Resolutions. All metrics are calculated at the original image resolution of the samples in iHarmony4. The inputs to the existing methods are in low-resolution. Their outputs are then bilinearly upsampled to high resolutions for metric calculation. We also apply Polynomial Color Mapping for upsampling (with subscript “+PCM”).

Dataset	Metric	DOVE[9]	DOVE[9] +PCM	BARG[8]	BARG[8] +PCM	IHT[12]	IHT[12] +PCM	Our
HAdobe5k	MSE↓	68.16	72.08	77.96	88.20	56.90	63.28	24.37
	fMSE↓	511.02	579.21	560.49	689.58	465.72	547.61	196.12
	PSNR↑	33.30	32.82	32.65	32.17	33.63	33.04	37.80
HFlickr	MSE↓	172.80	159.46	159.34	150.67	135.49	127.10	69.19
	fMSE↓	1192.88	1110.22	1114.29	1096.91	994.23	976.08	479.26
	PSNR↑	28.81	29.71	29.01	29.88	29.59	30.44	33.37
HCOCO	MSE↓	56.49	47.13	52.84	46.62	44.95	40.16	20.93
	fMSE↓	1000.14	844.84	940.79	844.21	838.86	785.03	374.96
	PSNR↑	33.35	34.5	33.54	34.51	34.19	34.85	37.69
Hday2night	MSE↓	58.23	67.81	53.99	66.91	63.26	72.94	37.28
	fMSE↓	1125.46	1007.99	958.02	1109.15	988.56	1054.97	640.74
	PSNR↑	35.44	35.17	35.65	35.07	35.71	35.06	37.15
All	MSE↓	72.98	67.35	71.93	70.76	58.89	57.22	27.62
	fMSE↓	882.26	800.48	851.83	832.62	750.06	741.98	339.23
	PSNR↑	32.86	33.49	32.82	33.32	33.54	33.83	37.23

Table 3. Comparison on Inference Speed, Model Size, and GPU Memory. The speed evaluation is conducted at 1080P resolution on a RTX3090 GPU.

Metric	S2AM[10]	DOVE[9]	BARG[8]	IntrIH[13]	IHT[12]	Our
Inference Speed (fps) ↑	6.76	13.8	11.6	1.2	5.1	56.3
Model Size (MB) ↓	268.1	219.1	234.9	163.5	25.8	21.7
GPU Memory (GB) ↓	6.3	6.5	3.7	16.47	18.5	2.3

the actual image harmonization performance. To study this issue, we further evaluate Harmonizer with the strong baseline DOVE and the state-of-the-art BARG/IHT at high resolutions. We compute the metrics at the original resolutions of the images in iHarmony4. Note that the subsets in iHarmony4 have different resolutions, *e.g.*, the average image size for HCOCO is about 500×500 and for HAdobe5k is about 3000×3000 . Since high-resolution inputs would significantly degrade the performances of the existing methods that are trained on 512×512 resolution, we still input low-resolution images to them and then bilinearly upsample their results to high resolutions for metric calculation. To avoid blurry outputs caused by bilinear upsampling, we also upsample their results using Polynomial Color Mapping (PCM) [1], which can transfer the foreground appearances in the low-resolution outputs to the high-resolution composite images without loss of details (has been used for visualization in previous works). In contrast, our Harmonizer can process the high-resolution inputs directly as its filters are resolution-independent. Table 2 shows the metrics computed at the original image resolutions. The performances of the existing methods are significantly lowered, *e.g.*, MSE/fMSE of IHT are increased from 37.07/395.66 to

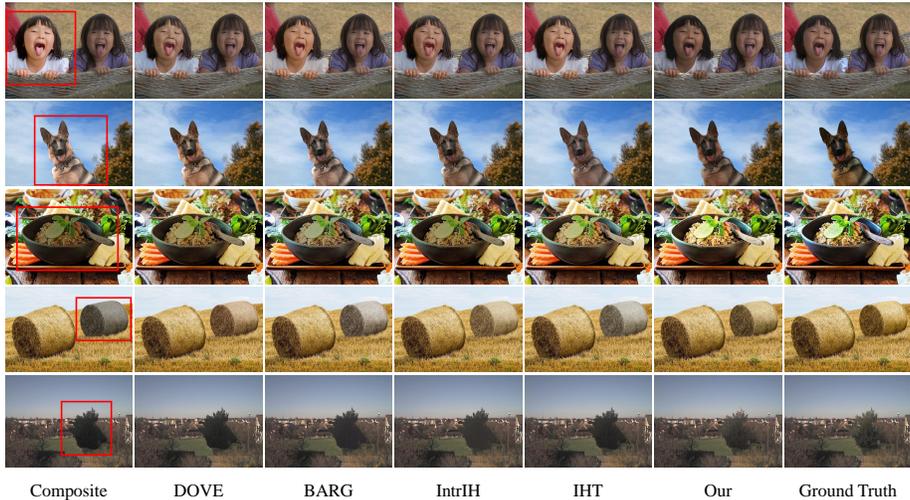


Fig. 6. Visual Comparison on iHarmony4. The red boxes in the composite images indicate the foreground regions. Zoom in for better visualization.

57.22/741.98. In contrast, Harmonizer only has a small performance drop at high resolutions, and its MSE is now 50% lower than the state-of-the-art method. We provide some visual comparisons in Fig. 6.

Table 3 compares different methods in terms of inference speed, model size, and memory requirement. A fast inference speed is necessary for real-time applications, while a small model size and a low memory requirement facilitate deployment on mobile devices. Our results demonstrate that Harmonizer is faster, lighter, and more memory efficient than other methods. Remarkably, on a RTX3090 GPU, Harmonizer can process a 1080P (Full HD) video at 56 *fps*, about 4× faster than the recent fastest method DOVE [9]. Moreover, we observe that Harmonizer can be further accelerated by a fusion implementation of filters or using techniques like Halide Auto-Scheduler.

4.2 Video Harmonization Results

As shown in Fig. 7, by applying Eq. 14, Harmonizer obtains stable harmonization results across video frames. On the contrary, the results of the prior methods suffer from severe flickers. Unfortunately, our strategy proposed in Sec. 3.4 is not suitable for use in prior existing methods since they solve harmonization as a pixel-wise image translation problem in a black-box manner.

4.3 Ablation Study

In Table 4 *Left*, we evaluate the cascade regressor and dynamic loss strategy proposed in Harmonizer. The results show that both techniques can improve the image harmonization performance. We also observe that even without these two

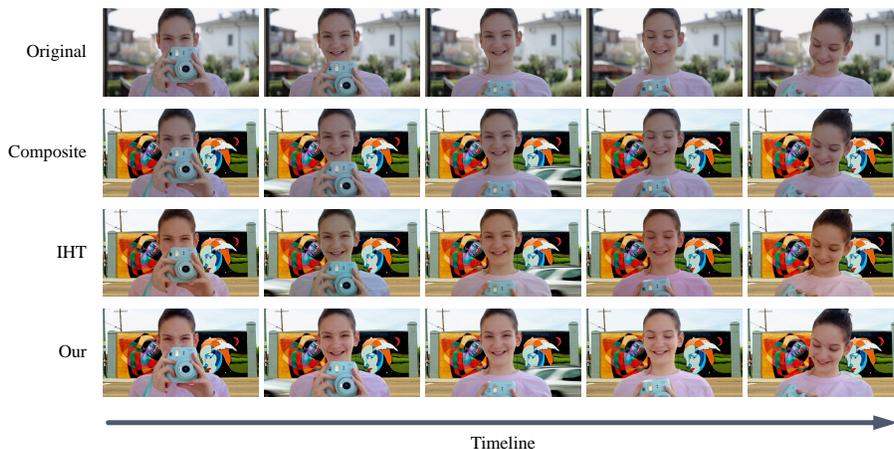


Fig. 7. Video Harmonization Results. Previous image harmonization methods (we only visualize the results from IHT [12] here due to space limitation) output the frames with inconsistent foreground appearances (*e.g.*, the cloth regions). Instead, our Harmonizer obtains consistent harmonization results across frames.

Table 4. Ablation of Harmonizer on iHarmony4. *Left:* Evaluating the effectiveness of different Harmonizer components at 256x256 resolution. *Right:* Evaluating the performance of Harmonizer with different numbers of filters. MSE and PSNR are calculated at 256x256 resolution, while *fps* is measured at 1080P resolution.

Cascade Regressor	Dynamic Loss Strategy	Metrics		Metrics	Number of Filters			
		MSE ↓	PSNR ↑		2	4	6	8
		28.47	36.96	MSE ↓	74.16	29.60	24.26	23.51
✓		26.85	37.23	PSNR ↑	33.49	36.75	37.84	38.06
✓	✓	24.26	37.84	<i>fps</i> ↑	86.2	63.5	56.3	51.9

techniques, the results of Harmonizer (MSE of 33.28) still surpass the previous state-of-the-art (MSE of 37.07), which demonstrates that the overall design of Harmonizer has advantages over the black-box autoencoders. In Table 4 *Right*, we analyze the impact of different numbers of filters on the performance and speed. Specifically, we validate Harmonizer with 2, 4, 6, and 8 filters. For the experiments with 6 or 8 filters, we choose the most important filters based on Fig. 2. For the experiments with 2 or 4 filters, we select the filters randomly and report the metrics averaged over 3 runs. The results show that the performance of Harmonizer increases as the number of filters increases. Besides, Harmonizer needs at least 4 filters to avoid large performance degradation.

4.4 User Studies

We show the advantage of Harmonizer in real-world image/video harmonization through user studies. For image harmonization, we use the real composite images released in [39], which includes 99 images. Since these images have no labels,

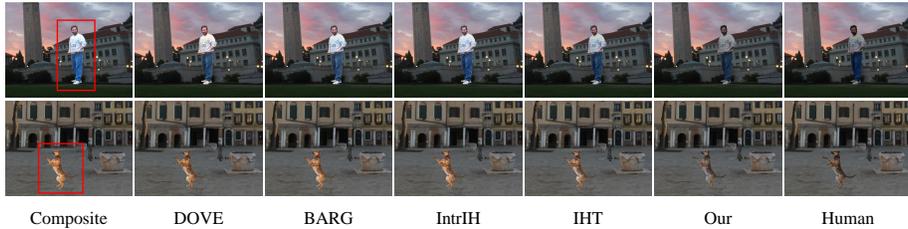


Fig. 8. Visual Comparison on Real Composite Images. “Human” means that the results are produced by humans.

Table 5. User Study Results. We calculate B-T scores to quantify our user study results. For image harmonization, the results from humans are compared.

Metrics	Composite	DOVE[9]	BARG[8]	IntrIH[13]	IHT[12]	Our	Human
B-T Score (Image) \uparrow	0.412	0.639	0.618	0.663	0.724	1.028	1.393
B-T Score (Video) \uparrow	1.173	0.587	0.497	0.530	0.601	2.042	-

we ask a skilled human to process them for reference. For video harmonization, we composite the 20 foreground videos generated by the video matting methods [32,21,26] with 20 new background videos to create composite videos (one of them is shown in Fig. 7). We invite 12 participants to rank the results from different methods and the human. In Table 5, we follow prior works to use the Bradley-Terry model (B-T model) [6] for ranking. Harmonizer achieves the highest B-T scores. The B-T scores of applying prior methods to video harmonization are even lower than the original composite inputs due to severe flickering. Fig. 8 visualizes two samples used in our image harmonization user study.

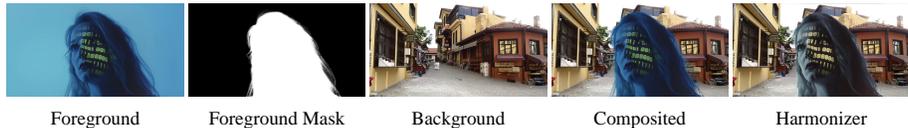


Fig. 9. A Failure Case of Harmonizer. We show a composite image with a strong hue difference (within the blue color range) between foreground and background.

5 Conclusion

In this paper, we have studied the image harmonization process carried out by humans, which has inspired us to design Harmonizer. Benefited by our novel architecture and two proposed techniques, *i.e.*, a cascade regressor and a dynamic loss strategy, Harmonizer is lighter and faster than prior methods while achieving new state-of-the-art performances. Nevertheless, our method does have limitations. It may fail to handle color-specific appearance inconsistency or the different lighting conditions between the foreground and background. Fig. 9 shows one case. As a future work, we would like to develop more complex image filters, *e.g.*, color-separated filters, to address the problem.

References

1. Affi, M., Brown, M.S.: Deep white-balance editing. In: CVPR (2020)
2. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics* (2009)
3. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. *IEEE TPAMI* (2014)
4. Bonneel, N., Sunkavalli, K., Paris, S., Pfister, H.: Example-based video color grading (2013)
5. Bonneel, N., Tompkin, J., Sunkavalli, K., Sun, D., Paris, S., Pfister, H.: Blind video temporal consistency. *ACM Transactions on Graphics* (2015)
6. Bradley, R.A., Terry, M.E.: The rank analysis of incomplete block designs — I. The method of paired comparisons. *Biometrika* (1952)
7. Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.Q.: Color harmonization. *ACM Transactions on Graphics* (2006)
8. Cong, W., Niu, L., Zhang, J., Liang, J., Zhang, L.: Bargainnet: Background-guided domain translation for image harmonization. In: ICME (2021)
9. Cong, W., Zhang, J., Niu, L., Liu, L., Ling, Z., Li, W., Zhang, L.: Dovenet: Deep image harmonization via domain verification. In: CVPR (2020)
10. Cun, X., Pun, C.M.: Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Process* (2020)
11. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: *NeurIPS* (2014)
12. Guo, Z., Guo, D., Zheng, H., Gu, Z., Zheng, B., Dong, J.: Image harmonization with transformer. In: *ICCV* (2021)
13. Guo, Z., Zheng, H., Jiang, Y., Gu, Z., Zheng, B.: Intrinsic image harmonization. In: *CVPR* (2021)
14. Hao, G., Iizuka, S., Fukui, K.: Image harmonization with attention-based deep feature modulation. In: *BMVC* (2020)
15. Hu, Y., He, H., Xu, C., Wang, B., Lin, S.: Exposure: A white-box photo post-processing framework. *ACM Transactions on Graphics* (2018)
16. Huang, H., Xu, S., Cai, J., Liu, W., Hu, S.: Temporally coherent video harmonization using adversarial networks. *IEEE Transactions on Image Process* (2020)
17. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *CVPR* (2017)
18. Jia, J., Sun, J., Tang, C.K., Shum, H.Y.: Drag-and-drop pasting. *ACM Transactions on Graphics* (2006)
19. Jiang, Y., Zhang, H., Zhang, J., Wang, Y., Lin, Z., Sunkavalli, K., Chen, S., Amirghodsi, S., Kong, S., Wang, Z.: A self-supervised framework for image harmonization. In: *ICCV* (2021)
20. Johnson, M.K., Dale, K., Avidan, S., Pfister, H., Freeman, W.T., Matusik, W.: Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE Transactions on Visualization and Computer Graphics* (2010)
21. Ke, Z., Sun, J., Li, K., Yan, Q., Lau, R.W.: Modnet: Real-time trimap-free portrait matting via objective decomposition. In: *AAAI* (2022)
22. Lai, W.S., Huang, J.B., Wang, O., Shechtman, E., Yumer, E., Yang, M.H.: Learning blind video temporal consistency. In: *ECCV* (2018)

23. Lalonde, J.F., Efros, A.A.: Using color compatibility for assessing image realism. In: ICCV (2007)
24. Lei, C., Xing, Y., Chen, Q.: Blind video temporal consistency via deep video prior. In: Neurips (2020)
25. Liang, J., Cun, X., Pun, C.: Spatial-separated curve rendering network for efficient and high-resolution image harmonization. Arxiv **abs/2109.05750** (2021)
26. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. In: WACV (2022)
27. Ling, J., Xue, H., Song, L., Xie, R., Gu, X.: Region-aware adaptive instance normalization for image harmonization. In: CVPR (2021)
28. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep painterly harmonization. EGSR (2018)
29. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Transactions on Graphics (2003)
30. Pitie, F., Kokaram, A.: The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In: European Conference on Visual Media Production (2007)
31. Pitie, F., Kokaram, A., Dahyot, R.: N-dimensional probability density function transfer and its application to color. In: ICCV (2015)
32. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. vol. 106, p. 107404 (2020)
33. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications (2001)
34. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
35. Sofiuk, K., Popenova, P., Konushin, A.: Foreground-aware semantic representations for image harmonization. In: WACV (2021)
36. Sunkavalli, K., Johnson, M.K., Matusik, W., Pfister, H.: Multi-scale image harmonization. ACM Transactions on Graphics (2010)
37. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: ICML (2019)
38. Tao, M.W., Johnson, M.K., Paris, S.: Error-tolerant image compositing. In: ECCV (2010)
39. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: CVPR (2017)
40. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
41. Wang, X., Yu, J.: Learning to cartoonize using white-box cartoon representations. In: CVPR (2020)
42. Xue, S., Agarwala, A., Dorsey, J., Rushmeier, H.: Understanding and improving the realism of image composites. ACM Transactions on Graphics (2012)
43. Yan, Z., Zhang, H., Wang, B., Paris, S., Yu, Y.: Automatic photo adjustment using deep neural networks. ACM Transactions on Graphics (2016)
44. Zaragoza, J., Chin, T.J., Brown, M.S., Suter, D.: As-projective-as-possible image stitching with moving dlt. In: CVPR (2013)
45. Zhang, F., Liu, F.: Parallax-tolerant image stitching. In: CVPR (2014)
46. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape-from-shading: a survey. IEEE TPAMI (1999)

47. Zhu, J.Y., Krahenbuhl, P., Shechtman, E., Efros, A.A.: Learning a discriminative model for the perception of realism in composite images. In: ICCV (2015)
48. Zou, Z., Shi, T., Qiu, S., Yuan, Y., Shi, Z.: Stylized neural painting. In: CVPR (2021)