Semi-supervised Parametric Real-world Image Harmonization

Ke Wang[1,2], Michaël Gharbi[1], He Zhang[1], Zhihao Xia[1], Eli Shechtman[1]
[1] Adobe Inc.
[2] EECS, University of California, Berkeley
{kewang, mgharbi, hezhan, zxia, elishe}@adobe.com
kewang@berkeley.edu

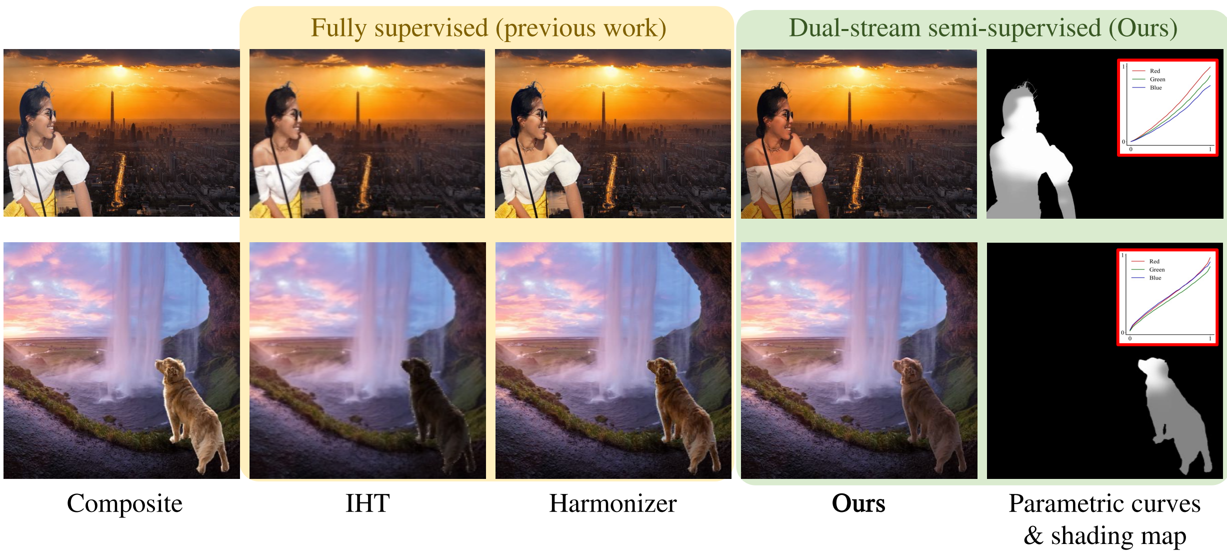

Figure 1. **Visual comparisons between state-of-the-art harmonization methods IHT [9], Harmonizer [14], and ours.** Our model is fully parametric. This gives artists full posterior control over the final composite, makes runtime efficient for high-resolution real-world inputs and regularizes training. Our model predicts *global RGB curves* and a *local shading map* (right). Benefiting from the novel dual-stream semi-supervised training strategy, our method (right) produces more realistic harmonized images on real-world composites (left). This new training strategy, together with the shading map, makes it the first harmonization method to address local tonal adjustments, such as shading the face according to the sun's direction (top) or selectively darkening the part of the dog inside the cave (bottom).

Abstract

Learning-based image harmonization techniques are usually trained to undo synthetic random global transformations applied to a masked foreground in a single ground truth photo. This simulated data does not model many of the important appearance mismatches (illumination, object boundaries, etc.) between foreground and background in real composites, leading to models that do not generalize well and cannot model complex local changes. We propose a new semi-supervised training strategy that addresses this problem and lets us learn complex local appearance harmonization from unpaired real composites, where foreground and background come from different images. Our model is fully parametric. It uses RGB curves to correct the global colors and tone and a shading map to model local variations. Our method outperforms previous work on established benchmarks and real composites, as shown in a user study, and processes high-resolution images interactively.

Code, and project page available at: https://kewang0622.github.io/sprih/.

1. Introduction

Image harmonization [12, 22, 23, 26, 28, 32] aims to iron out visual inconsistencies created when compositing a foreground subject onto a background image that was captured under different conditions [18, 32], by altering the foreground's colors, tone, etc., to make the composite more realistic. Despite significant progress, the practicality of today's most sophisticated learning-based image harmonization techniques [3, 4, 9, 10, 13, 14, 16, 32] is limited by a severe domain gap between the synthetic data they are trained on and real-world composites.

As shown in Figure 2, the standard approach to generating synthetic training composites applies global transforms

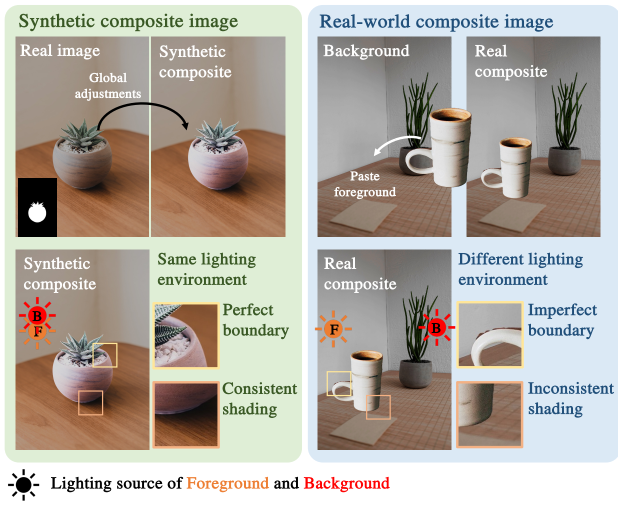

Figure 2. **Domain Gap between synthetic and real-world composites.** The existing synthetic composites [4] (left), generated by applying global transforms (e.g., color, brightness), are unable to simulate many of the appearance mismatches that occur in real composites (right). This leads to a domain gap: models trained on synthetic data do not generalize well to real composites. In real composites (right), the foreground and background are captured under different conditions. They have different illuminations, the shadows do not match, and the object's boundary is inconsistent. Such mismatches do not happen in the synthetic case (left).

(color, brightness, contrast, etc.) to a masked foreground subject in a ground truth photo. This is how the iHarmony Dataset [2, 4] was constructed. A harmonization network is then trained to recover the ground truth image from the synthetic input. While this approach makes supervised training possible, it is unsatisfying in simulating the real composite in that synthetic data does not simulate mismatch in illumination, shadows, shading, contacts, perspective, boundaries, and low-level image statistics like noise, lens blur, etc. However, in real-world composites, the foreground subject and the background are captured under different conditions, which can have more diverse and arbitrary differences in any aspects mentioned above.

We argue that using realistic composites for training is essential for image harmonization to generalize better to real-world use cases. Because collecting a large dataset of artist-created before/after real composite pairs would be costly and cumbersome, our strategy is to use a semi-supervised approach instead. We propose a novel dual-stream training scheme that alternates between two data streams. Similar to previous work, the first is a supervised training stream, but crucially, it uses artist-retouched image pairs. Different from previous datasets, these artistic adjustments include global color editing but also dodge and burn shading corrections and other local edits.

The second stream is fully unsupervised. It uses a GAN [8] training procedure, in which the critic compares our harmonized results with a large dataset of realistic image composites. Adversarial training requires no paired ground truth. The foreground and background for the composite in this dataset are extracted from different images so that their appearance mismatch is consistent with what the model would see at test time.

To reap the most benefits from our semi-supervised training, we also introduce a new model that is fully parametric. To process a high-resolution input composite at test time, our proposed network first creates a down-sampled copy of the image at 512×512 resolution, from which it predicts *global RGB curves* and a smooth, low-resolution *shading map*. We then apply the RGB curves pointwise to the high-resolution input and multiply them by the upsampled shading map. The shading map enables more realistic local tonal variations, unlike previous harmonization methods limited to global tone and color changes, either by construction [14, 16, 31] or because of their training data [4].

Our parametric approach offers several benefits. First, by restricting the model's output space, it regularizes the adversarial training. Unrestricted GAN generators often create spurious image artifacts or other unrealistic patterns [36]. Second, it exposes intuitive controls for an artist to adjust and customize the harmonization result post-hoc. This is unlike the black-box nature of most current learning-based approaches [3, 4, 9, 10], which output an image directly. And, third our parametric model runs at an interactive rate, even on very high-resolution images (e.g., 4k), whereas several state-of-the-art methods [4, 9, 10] are limited to low-resolution (e.g., 256×256) inputs.

To summarize, we make the following contributions:

- A novel dual-stream semi-supervised training strategy that, for the first time, enables training from real composites, which contains much richer local appearance mismatches between foreground and background.
- A parametric harmonization method that can capture these more complex, local effects (using our shading map) and produces more diverse and photorealistic harmonization results.
- State-of-the-art results on both synthetic and real composite test sets in terms of quantitative results and visual comparisons, together with a new evaluation benchmark.

2. Related works

Image harmonization. Traditional image harmonization methods mainly focus on adjusting the low-level appearance statistics (e.g., color statistics, gradient information) between the foreground objects and the background [12, 22, 23, 26, 28, 32]. Supervised learning-based approaches have been proposed and shown notable success [3, 4, 9, 10, 29, 37] by learning image harmonization from synthetic training pairs, for instance, iHarmony Dataset [4]. Works as DIH

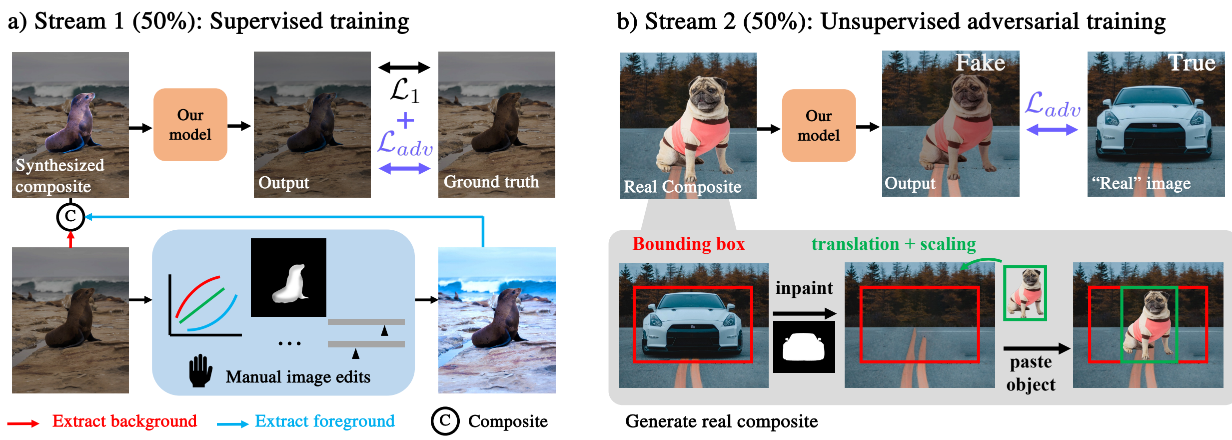

Figure 3. **Overview of semi-supervised dual-stream training strategy.** To bridge the domain gap, our proposed semi-supervised dual-stream training strategy alternates between two training streams: a) Supervised training with artist-retouched composite image pairs (left). Artist adjustments include global color editing, shading correction, and other local edits. b) Unsupervised adversarial training with real-world composite images (right). It uses a GAN [8] training procedure, comparing our harmonized results with a large dataset of composite "real" images (see § 3.2 for details). The foreground and background for the composite are from different images, so the appearance mismatch is consistent with what we see at test time.

[29], DovNet [4], IHT [9], Guo *et al.* [10] consider the image harmonization task as a pixel-wise image-to-image translation task, and are limited to low-resolution inputs (typically 256×256) due to computational inefficiency. Recent work extended image harmonization to high-resolution images by designing parametric models [3, 14, 16, 31]. To name a few, Liang *et al.* learns the spatial-separated RGB curves for high-resolution image harmonization. Ke *et al.* [14] directly predicts the filter arguments of several white-box filters. In all of those approaches, synthetic training pairs are generated by applying global transforms to the masked foreground regions and hence do not simulate mismatch in illumination, shadows, shading, contact, etc., that happen in real-world composite images. Therefore, due to the synthetic training data and model construction [14, 16], previous works are limited to global tone and color changes. In contrast, our model is trained on real-world composite images and artist-retouched synthetic images, which enables us to model richer image edits and produce more compelling results on real composites.

Efficient and high-resolution image enhancement. There has been a wide range of research focusing on designing efficient and high-resolution image enhancement algorithms [6,7,17]. Gharbi *et al.* [6] introduced a convolutional neural network (CNN) that predicts the coefficients of a locally-affine model in bilateral space from down-sampled input images. The coefficients are then mapped back to the full-resolution image space. Zeng *et al.* [34] directly learns 3D Lookup Tables (LUTs) for real-time image enhancement. In our application, image harmonization can be considered as a background-guided image enhancement problem. Thus, inspired by [6, 34], we design a network that directly predicts the coefficients of RGB curves (piece-wise linear function) from down-sampled composite inputs. We then apply the RGB curves pointwise to the high-resolution input without introducing extra computation costs.

Image-based relighting Image-based relighting approaches [19, 21, 25, 33] focus on modifying the input lighting conditions and local shading to generate convincing composite results. However, recent relighting methods mainly focus on portraits and struggle to generalize to other objects, as Light-stage capture is limited to portraits and not diverse objects [5]. With a similar idea of incorporating local shading edits but a different approach, our method embeds the shading layer into a network and trains on composite image datasets without explicitly leveraging scene representations (geometry, materials, lighting) and using full relighting models.

3. Method

Our image harmonization method corrects the foreground subject in a rough composite to make the overall image look more realistic using a new parametric model (§ 3.1) that can be applied to real-world high-resolution images efficiently. Previous harmonization techniques train on synthetically-generated composite pairs [4], where the model's input is a global transformation of a ground truth image within a foreground subject mask. The colors are often unnatural, the mask boundary is close to perfect, and there is no mismatch in appearance, illumination, or low-level image statistics since both foreground and background come from the same image. As a result, models trained on such data generalize poorly. Our method addresses this crucial issue using a novel dual-stream semi-supervised train-

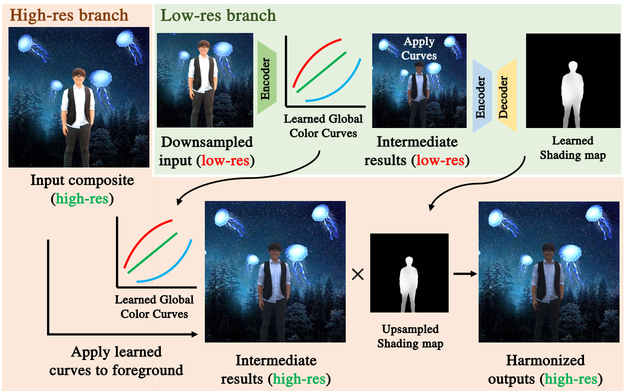

Figure 4. **Illustration of our parametric model design.** Our framework consists of a low-resolution branch and a high-resolution branch. At test time, we down-sample the given high-resolution image and predict the *global RGB curves and shading map* through a two-stage network. Those parametric outputs are then executed at the original resolution to produce the final harmonized image. Our model can scale to any resolution.

ing strategy (§ 3.2) that leverages high-quality artist-created image pairs and unpaired realistic composites to bridge the training-testing domain gap. See Figure 3 for an overview.

3.1. Parametric image harmonization

Our network design is inspired by real-world composite harmonization workflows[1]. An artist typically applies several image corrections sequentially, each dedicated to harmonizing a specific composite element, such as luminosity, color, or shading. Accordingly, our image transformation model consists of two modules, applied sequentially: a pointwise global RGB color correction module and a local shading correction using a low-frequency multiplicative map. For efficiency, our model operates at two resolutions.

Pipeline overview. As illustrated in Figure 4, our harmonization pipeline takes as input a foreground image $\mathbf{F} \in \mathbb{R}^{3hw}$ with dimensions $h, w \in \mathbb{N}$, a background image $\mathbf{B} \in \mathbb{R}^{3hw}$, and a compositing alpha mask $\mathbf{M} \in \mathbb{R}^{hw}$. We define the *unharmonized* composite image as $\mathbf{C} := \mathbf{M} \cdot \mathbf{F} + (1 - \mathbf{M}) \cdot \mathbf{B}$. At test time, we start by downsampling the inputs to a fixed resolution 512×512, denoting the low-resolution images by $\mathbf{C}^{lr}$, $\mathbf{B}^{lr}$, $\mathbf{M}^{lr}$ respectively. We concatenate these maps and pass them to a neural network f that predicts the parameters $[\theta_1, \theta_2] := f(\mathbf{C}^{lr}, \mathbf{B}^{lr}, \mathbf{M}^{lr})$ of our two-stage parametric image transformation. Finally, we apply the parametric transformation t_1, t_2 sequentially on the high-resolution input to obtain the final harmonized composite $\mathbf{O} := t_2(t_1(\mathbf{C}, \mathbf{M}; \theta_1), \mathbf{M}; \theta_2)$, where $\mathbf{M}$ is used to ensure only foreground mask area is altered. We describe the two stages in the parametric transformation next.

[1] https://youtu.be/g3qe4rDw1XU

Global color correction curves. In our first high-resolution processing stage t_1, we apply the predicted global RGB curves for color correction. These curves are parameterized as 3 piecewise linear curves with 32 control points and are applied independently to each color channel, resulting in a set of 2D coordinates, $\theta_1 \in \mathbb{R}^{32 \times 2 \times 3}$. The output color for each channel is interpolated between adjacent control points. We employ a ResNet-50-based network [11] to predict these parameters from $[\mathbf{C}^{lr}, \mathbf{B}^{lr}, \mathbf{M}^{lr}]$. The curve application, a per-pixel operation, allows for efficient computation at any resolution.

Local low-frequency shading map. Our second stage t_2 multiplies the image with a low-frequency grayscale shading map, to model local tonal corrections. It is applied to the output of the first stage. We constrain the shading map to only model low-frequency change by first generating θ_2 at a low resolution 64×64, then upsampling, and passing a single convolution layer at 512×512 to correct upsampling artifacts. It is produced by a modified U-Net [24] with large receptive field, given the low-resolution buffers $[\mathbf{C}^{lr}, \mathbf{B}^{lr}, \mathbf{M}^{lr}]$, together with the output of the color-correction stage at low-resolution $t_1(\mathbf{C}^{lr}, \mathbf{M}^{lr}; \theta_1)$. At test time, we upsample the low-resolution shading map to the original high-resolution and multiply it pointwise with the color-corrected image to obtain our final harmonized composite:

$$\mathbf{O} = t_1(\mathbf{C}, \mathbf{M}; \theta_1) \cdot \mathrm{upsample}(\theta_2). \tag{1}$$

3.2. Dual-stream semi-supervised training

Our semi-supervised training strategy aims to alleviate the generalization issues that plague many state-of-the-art harmonization models, as shown in Figure 2. During a *single training stage*, our approach equally samples *two data streams* and optimizes a distinct objective for each of them. The first stream uses input/output composite pairs similar to previous work, except that we only use artist-created image transformations instead of random augmentations. The second is unsupervised. This allows us to use more realistic images obtained by compositing foreground and background from unrelated images, for which no ground truth is easily obtainable. For the supervised stream, the objective combines ℓ_1 loss and adversarial loss, while the unsupervised stream solely utilizes adversarial loss.

Supervised training using retouched images. The first stream is fully supervised. Unlike previous work, we use images retouched by artists rather than mostly relying on random augmentations. We refer to this dataset as *Artist-Retouched* in the rest of the paper. Artists were allowed to use common image editing operations such as global luminosity or color adjustments, but also local editing tools like brushes, e.g., to alter the shading. Specifically, we collected $n = 46173$ before/after retouching image pairs

$\{\mathbf{I}_i, \mathbf{O}_i\}_{i=1,\dots,n}$, with the mask for one foreground object $\mathbf{M}_i$ for each pair. From each triplet, we can create 2 input composites for training: one with only the foreground retouched $\mathbf{M}_i \cdot \mathbf{O}_i + (1-\mathbf{M}_i) \cdot \mathbf{I}_i$, and the other with only the background is retouched $\mathbf{M}_i \cdot \mathbf{I}_i + (1-\mathbf{M}_i) \cdot \mathbf{O}_i$. Since our harmonization model only alters the foreground, we use the unedited image $\mathbf{I}_i$, and the retouched image $\mathbf{O}_i$ as ground truth targets for these input composites, respectively.

When sampling training data from this stream, we optimize our model's parameters to minimize the sum of an ℓ_1 reconstruction error $\mathcal{L}_{rec}$ between the ground truth and our model output, and an adversarial objective [8]

$$\lambda \mathcal{L}_{rec} + (1-\lambda)\mathcal{L}_{\mathrm{G}}, \tag{2}$$

with λ balances the two losses. For our experiments, λ is empirically set to 0.92. The generator, our parametric image harmonization model, is trained to produce outputs that cannot be distinguished from "real" images. We use a U-Net discriminator [30] D to make per-pixel real vs. fake classifications. Since our data formation model assumes the background is always correct, our discriminator is trained to predict the inverted foreground mask $1 - \mathbf{M}$. That is when shown "fake" images, i.e., the background pixels have label 1 and the foreground 0. For the "real" class, the target is all an all-1s map. So the discriminator loss is given by:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{\mathbf{I}_{real}}[\log(D(\mathbf{I}_{real}))] \\ &-\mathbb{E}_{\mathbf{I}_{fake}}[\log((1-\mathbf{M}) - D(\mathbf{I}_{fake}))],\end{aligned} \tag{3}$$

The generator loss is:

$$\mathcal{L}_G = -\mathbb{E}_{\mathbf{I}_{fake}}[\log(D(\mathbf{I}_{fake}))]. \tag{4}$$

To further increase the training diversity, we randomly augment the foreground brightness on the fly without retouching the color.

Unsupervised training with real composites. Our second training stream is unsupervised. It uses randomly generated composites that are representative of real-world use cases but for which no ground truth is available. To properly reproduce the appearance mismatch in real applications, we create these composites as follows. We start from a dataset of m images $\{\mathbf{I}_i\}_{i=1,\dots,m}$, each with a foreground object mask $\mathbf{M}_i$, from which we derive a foreground $\mathbf{F}_i = \mathbf{M}_i \cdot \mathbf{I}_i$ and a background $\mathbf{B}_i = (1-\mathbf{M}_i) \cdot \mathbf{I}_i$. As preprocessing, we dilate the foreground mask by 30 pixels and inpaint the corresponding area in the background image using a pre-trained inpainting network (we use LaMa [27]). Then during, training we sample two images i and j and create a composite by pasting the foreground j onto the inpainted background of i:

$$\mathbf{C}_{ij} := \mathbf{F}_j \cdot \mathbf{M}_j + \text{inpaint}(\mathbf{B}_i, \mathbf{M}_i) \cdot (1-\mathbf{M}_j). \tag{5}$$

The triplet $[\mathbf{C}_{ij}, \text{inpaint}(\mathbf{B}_i, \mathbf{M}_i), \mathbf{M}_j]$ is passed as input to our model. Figure 3b illustrates the process. $\mathbf{F}_j$ is translated and rotated from the original foreground so that it's maximally contained within $\mathbf{F}_i$'s bounding box.

With no ground truth available when sampling composites from this data stream, we only optimize the adversarial loss $(1-\lambda)\mathcal{L}_{\mathrm{G}}$, as defined in Eq. (4), where again the fake samples $\mathbf{I}_{fake}$ are the outputs of our model.

The discriminator is trained with Eq. (3), where $\mathbf{I}_{real}$ is not a real composite, but is obtained by masking the foreground subject $\mathbf{F}_i$, inpainting the background $\mathbf{B}_i$, and pasting the foreground back onto the same image, i.e.

$$\mathbf{I}_{real} := \mathbf{F}_i \cdot \mathbf{M}_i + \text{inpaint}(\mathbf{B}_i, \mathbf{M}_i) \cdot (1-\mathbf{M}_i). \tag{6}$$

This is similar to how we produce a composite of two images i and j, expect that we only use one image, i. This alteration of the "real" class is to prevent the discriminator from using the inpainting boundary region as a strong cue to discriminate between our model output and real images, which leads to collapse in the GAN training.

GAN training is known to be unstable or cause image artifacts [36], but because our parametric harmonization model adjusts color curves and adds low-resolution shadows, instead of predicting pixels directly, it has a strong regularizing effect, which prevents the GAN training to degenerate and cause spurious artifacts in the output image. We use the same discriminator (and generator) in both streams.

4. Experiments

We compare our parametric image harmonization model with state-of-the-art methods on established benchmarks (§ 4.1), as well as a test subset of *Artist-Retouched* dataset. Furthermore, we demonstrate our superior performance on real-world harmonization tasks via a user study and qualitative comparisons on real composites (§ 4.2). Ablation studies highlight the advantages of our semi-supervised training approach and our parametric model's components (§ 4.3). More results can be found in the supplementary.

Evaluation metrics: For quantitative comparisons with ground truth, we report performances by Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [35]. PSNR is measured in dB and calculated as: $\text{PSNR} = 10\log_{10}\frac{255^2}{\text{MSE}}$.

Implementation details: Our models are implemented in PyTorch [20] and trained on an NVIDIA A100 GPU using the Adam optimizer [15] for 80 epochs, with a batch size of 8 and an initial learning rate of 4×10^{-5}, decayed by a factor 0.2 every 20 epochs. Our model has 93M parameters (23M for stage t_1, 70M for stage t_2). Our model can run at an interactive rate where inference at 512×512 resolution takes on average (100 independent runs) 377 ms on an Apple M1 CPU, and 48.6 ms on an NVIDIA A100 GPU.

4.1. Quantitative comparisons on paired data

We compare our method with three recent methods, DovNet [4], Image Harmonization with Transformer

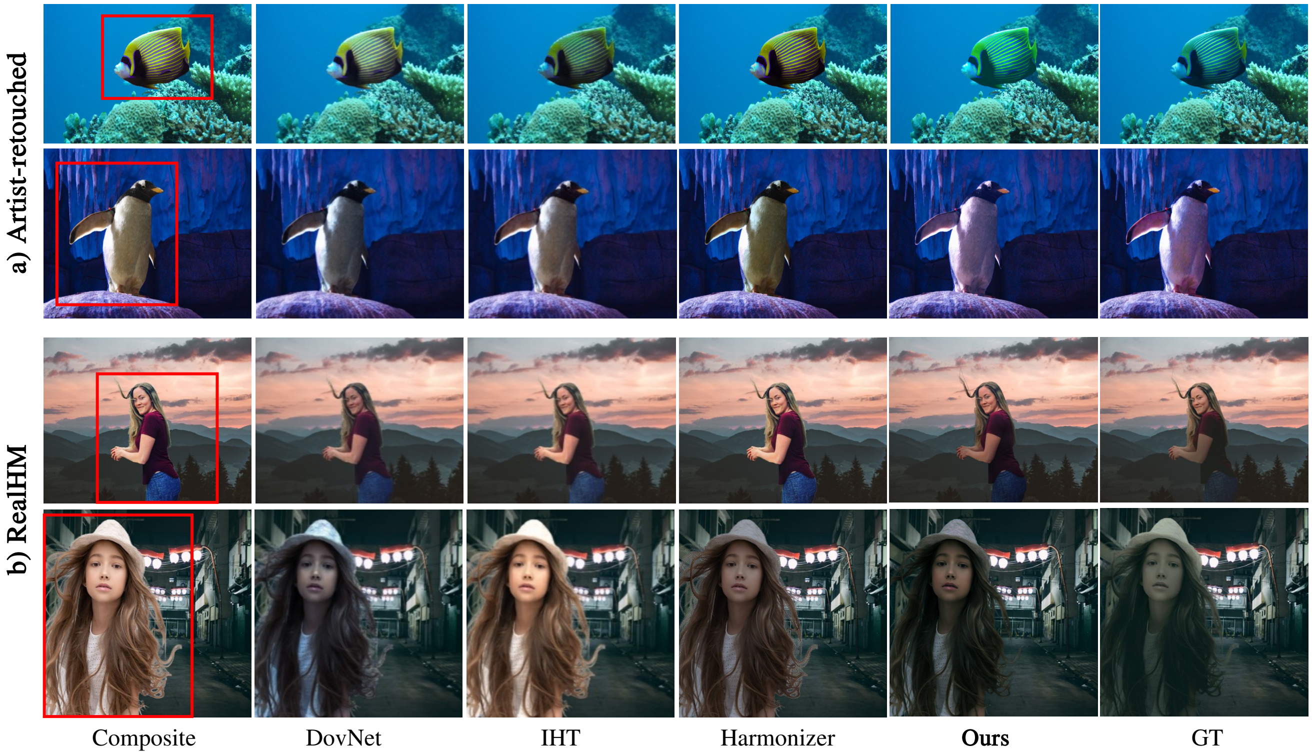

Figure 5. **Representative visual comparisons between state-of-the-art harmonization results.** We compared our method with composite, DovNet [4], IHT [9], and Harmonizer [14], and ground truth on both a) *Artist-Retouched* synthetic dataset and b) RealHM real-world composite dataset. Red boxes indicate the foreground subject in the composite image. The ground truth for RealHM benchmark [13] is expert-annotated harmonization results. Our results show better visual agreements with the ground truth in terms of color harmonization (rows 1,2 and 4) and shading correction (row 3).

(IHT) [14], and Harmonizer [14], using the pre-trained model released publicly by the authors. We first evaluate the synthetic iHarmony benchmark [4]. For fairness, our method uses the same setup as theirs for this comparison. In particular, we train our model exclusively on the same training set as the baselines, using only our fully-supervised stream, deactivating the adversarial loss, and only passing the composite $\mathbf{C}$ and foreground mask $\mathbf{M}$ as inputs. We report metrics at both at 256×256 resolution and at 2048×2048 on the HAdobe5k high-resolution subset of iHarmony. Like our parametric approach, Harmonizer can process high-res images, but the other two methods are limited to 256×256 inputs. So, for high-res comparison, we bilinearly downsample the input to DovNet and IHT, process the image, then bilinearly upsample the result before computing the metrics. Despite its simplicity, our parametric model consistently outperforms or matches the more complex baselines. Results are summarized in Table 1.

The iHarmony dataset is dominated by unrealistic synthetic image augmentations (71%), so we also evaluate our results on more realistic retouches from human experts. The two datasets we use for evaluation are a testing split of our *Artist-Retouched* dataset, introduced in Section 3.2, containing 1000 before/after pairs, and the RealHM [13] benchmark, containing 216 real-world high-resolution compos-

Size	Method	MSE ↓	PSNR ↑	SSIM ↑ $\times10^{-2}$	LPIPS ↓ $\times10^{-3}$
256	Composite	172.3	31.74	97.48	16.46
	DovNet [4]	51.33	34.97	98.12	9.734
	IHT [9]	30.46	37.33	98.77	**7.347**
	Harmonizer [14]	**24.24**	**38.25**	**99.09**	7.349
	Ours	**20.57**	**38.30**	**98.91**	**7.270**
2048*	Composite	352.9	28.39	96.36	14.52
	DovNet [4]	66.37	34.01	96.35	21.45
	IHT [9]	47.34	35.12	96.53	20.65
	Harmonizer [14]	**23.30**	**38.33**	**98.77**	**7.148**
	Ours	**20.31**	**38.29**	**98.82**	**7.123**

Table 1. **Quantitative comparison on iHarmony benchmark [4]** at both 256×256 and 2048×2048. ($*$) We only calculate the metrics on the Adobe5k dataset (a subset of iHarmony4) for high-resolution images. **Red**, and **Blue** correspond to the first and second best results. ↑ means higher the better, and ↓ means lower the better.

ites with expert annotated harmonization results as ground truth. We compared the performance of their pre-trained models and ours trained with the full dual-stream pipeline at 2048×2048 resolution. Table 2 shows our method consistently outperforms the baselines, with around 30% relative MSE improvements compared to Harmonizer [14] on both

Dataset	Method	MSE ↓	PSNR ↑	SSIM ↑ $\times 10^{-2}$	LPIPS ↓ $\times 10^{-3}$
Artist-Retouched	Composite	603.20	23.41	91.19	40.18
	DovNet [4]	352.4	26.42	90.83	56.47
	IHT [9]	369.3	26.36	90.87	55.80
	Harmonizer [14]	**239.1**	**29.42**	**93.84**	**33.75**
	Ours	**170.1**	**29.79**	**94.56**	**29.18**
RealHM	Composite	404.4	25.88	94.70	29.32
	DovNet [4]	225.1	26.72	92.00	47.50
	IHT [9]	264.0	26.48	92.46	48.48
	Harmonizer [14]	**231.4**	**27.40**	**94.86**	**27.62**
	Ours	**153.3**	**28.34**	**95.51**	**23.09**

Table 2. **Quantitative Comparison on RealHM benchmark and *Artist-Retouched* dataset.** Our approach outperforms other methods in all four metrics.

datasets. As shown in Figure 5, our method produces more realistic results, closer to the ground truth.

4.2. Evaluation on real composite images

Our semi-supervised training procedure allows us to train on realistic composites, where foreground and background come from different sources. Just like it limits the training potential of harmonization methods, using paired data created from a single ground truth image for evaluation is unsatisfying because it is not representative of real-world use cases (Fig. 2). So, we demonstrate the practical effectiveness of our method in a user study with real composites. For qualitative evaluation, we also created a set of 40 high-resolution real composite images with reference images.

User Study. Our user study follows a 2 alternatives forced choice protocol [35], comparing our model with DovNet [4], IHT [9], and Harmonizer [14]. We selected 60 real composites from the RealHM dataset [13], making sure there were no duplicate foregrounds or backgrounds. Since RealHM primarily focuses on portrait images, we also created 40 non-portrait real composites using free-to-use images from Unsplash [2], giving us a total of 100 real composite images. Each of our results is compared with the unaltered input composite and the three baseline results, which gives $100 \times 4 = 400$ image pairs to compare in total, which we submitted for evaluation to a pool of subjects on Amazon Turk [3]. Each participant was shown 50 image pairs and, for each pair, they were asked to "select which image looks more plausible". To ensure the quality of the responses, each subject was also shown 10 'sentinel' testing pairs composed of a real natural image and an extremely off-retouch image (e.g., where the image is all green). This helped us filter low-quality participants, such as users that always click 'left' to try and game the MTurk reward. After filtering, we obtained pair-wise comparison results from 70 subjects, contributing a total of 3500 comparisons. To analyze these results, we follow previous work [3,4,14], and use the Bradley-Terry (B-T) [1] model to derive the global ranking of all methods. We normalize the B-T scores such that the sum of the scores equals one across methods. Table 3 summarizes the results. It shows that our method achieves the highest B-T score, outperforming all the baselines, indicating our approach compares favorably in real-world image harmonization.

[2] https://unsplash.com/
[3] https://www.mturk.com/

Methods	B-T Score ↑
Composite	0.1025
DovNet [4]	0.1342
IHT [9]	0.2350
Harmonizer [14]	0.2257
Ours	**0.3025**

Table 3. **User Study Results.** B-T scores of composite image, DovNet [4], IHT [9], Harmonizer [14] are calculated on 100 real composite images. Our approach ranks first, suggesting superior real-world performance.

Real composites with captured reference. Figure 6 shows two representative examples of real composite results (see supplemental for more). For this qualitative comparison, we created a dataset of 40 high-resolution real-composite images with reference images by capturing a fixed set of foreground objects against multiple backgrounds, as well as a 'background-only' image. By segmenting the foreground object from one photo and pasting onto the 'background-only' image of another, we get an input composite for our model. The captured photo of the same object in the same background scene (placed at roughly the same location) acts as qualitative reference. Compared to other approaches, our results are visually closer to the captured reference.

4.3. Ablation studies

We assess the advantages of our semi-supervised dual-stream training approach, contrasting it with traditional supervised training, while also examining the effects of our global RGB curve module and shading map. We conduct the comparisons on RealHM [13] at 2048×2048, comparing our full method (dual-stream training + two-stage model) with: 1. Supervised training only (Stream 1) + global curves only; 2. Supervised training only (Stream 1) + two-stage parametric model; 3. Dual-stream training + global curves module only. We report quantitative metrics (MSE and PSNR), and the B-T score from a user study (similar to § 4.2, but with 68 subjects). Table 4 and Figure 7 summarize our findings, revealing that our shading map and dual-stream training strategy substantially enhance realism

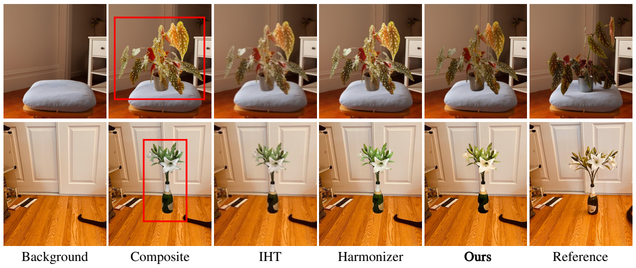

Figure 6. **Real composite harmonization results with captured reference.** The composite is obtained by pasting the foreground subject, from a different photo (not shown) onto the background (left). The reference (right) is obtained by physically placing the foreground subject in the background scene and taking a photo. We compare our method with IHT [9], and Harmonizer [14]. Our results show better visual agreement with the captured reference (best viewed by zooming on the digital preprint).

compared to the curve-only, fully-supervised model.

Stream 1	Stream 2	Global Curves	Shading Map	MSE	PSNR	B-T score
✓		✓		291.4	26.32	0.201
✓		✓	✓	268.3	26.60	0.206
✓	✓	✓		223.8	27.23	0.217
✓	✓	✓	✓	**153.3**	**28.34**	**0.252**
Composite				-	-	0.124

Table 4. **Ablation study results of training strategies and parametric model.** We compare our semi-supervised training strategy (Stream 1 + Stream 2) with supervised training (Stream 1) and compare our two-stage model (Global Curves + Shading map) versus the model with only the global curve module. MSE and PSNR are used for quantitative comparisons, and the B-T score is calculated from user study results.

As reported in Table 4, we observe that the dual-stream training strategy outperforms supervised training (row 3 and 4 *v.s.* row 1 and 2) in terms of both quantitative metrics and B-T score, which demonstrates the benefits of our proposed dual-training strategy in real-world applications. Inspecting the results in Figure 7, we observe that the dual-training strategy (column 4 and 5) brings advantages in color-harmonization when there is a strong foreground-background color mismatch.

On the other hand, as shown in Table 4 row 3 *v.s.* row 4, our proposed two-stage parametric model outperforms the global curve-only model by a large margin on RealHM benchmark, reducing the MSE by 30%. Furthermore, as shown in Figure 7, our full model (last column) includes both color harmonization and local shading to the results, achieving more plausible and harmonious results.

To better visualize the roles of our two-stage parametric model, Figure 8 shows the intermediate results as well as the parametric outputs (global curves and shading map) of a representative example. The global curves module harmonizes the global tone of the foreground sculpture and matches it with the background scene, while the shading map module refines local adjustments to harmonize the sculpture's shading with the lighting environment.

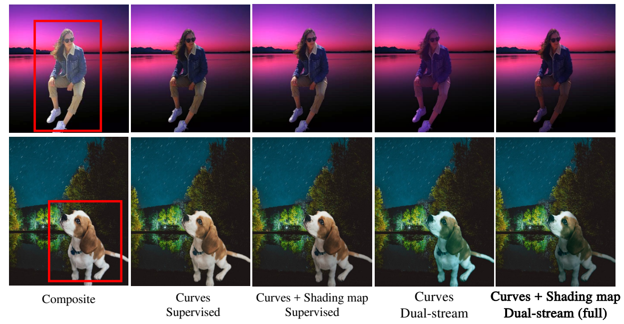

Figure 7. **Visual comparison of ablations.** Our full pipeline (right) shows more color-harmonious results than supervised training-only models (columns 2 and 3). Our local shading map adjusts local shading and produces more natural outputs (compare columns 4 and 5).

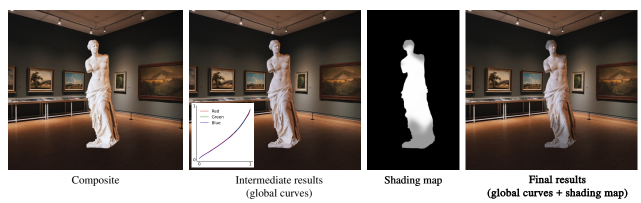

Figure 8. **Intermediate results and parametric outputs.** RGB curves harmonize the global color/tone (center), while our shading map corrects the local shading in the harmonization output (right).

5. Conclusion

In this work, we propose a novel semi-supervised dual-stream training strategy to bridge the training-testing domain gap and mitigate the generalization issues that limit previous works for real-world image harmonization. Our method leverages high-quality artist-created image pairs and unpaired realistic composites to enable richer image edits for real-world applications. Besides, we introduce a new two-stage parametric model (*Global RGB Curves* and *shading map*) to reap the most benefits from our training strategy and, for the first time, enable local editing effects with learned shading map. Our method outperforms other state-of-the-art methods on established benchmarks and real composites. Furthermore, our training strategy has the potential to generalize to a wider range of image harmonization operations (e.g., matching the noise, harmonizing the boundaries, adding cast shadows). As a future work, we would like to include more attributes in our models and further improve the performance of real-world image harmonization.

References

[1] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. 7

[2] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *The Twenty-Fourth IEEE Conference on Computer Vision and Pattern Recognition*, 2011. 2

[3] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18470–18479, 2022. 1, 2, 3, 7

[4] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8394–8403, 2020. 1, 2, 3, 5, 6, 7

[5] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 3

[6] Michaël Gharbi, Jiawen Chen, Jonathan T Barron, Samuel W Hasinoff, and Frédo Durand. Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 3

[7] Michaël Gharbi, YiChang Shih, Gaurav Chaurasia, Jonathan Ragan-Kelley, Sylvain Paris, and Frédo Durand. Transform recipes for efficient cloud photo enhancement. *ACM Transactions on Graphics (TOG)*, 34(6):1–12, 2015. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2, 3, 5

[9] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14870–14879, 2021. 1, 2, 3, 6, 7, 8

[10] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16367–16376, 2021. 1, 2, 3

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[12] Jiaya Jia, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Drag-and-drop pasting. *ACM Transactions on graphics (TOG)*, 25(3):631–637, 2006. 1, 2

[13] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4832–4841, 2021. 1, 6, 7

[14] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. *arXiv preprint arXiv:2207.01322*, 2022. 1, 2, 3, 6, 7, 8

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Jingtang Liang, Xiaodong Cun, and Chi-Man Pun. Spatial-separated curve rendering network for efficient and high-resolution image harmonization. *arXiv preprint arXiv:2109.05750*, 2021. 1, 2, 3

[17] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12826–12835, 2020. 3

[18] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021. 1

[19] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 3

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[21] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint indoor neural relighting from multi-view stereo. *ACM Transactions on Graphics (TOG)*, 40(5):1–18, 2021. 3

[22] Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1434–1439. IEEE, 2005. 1, 2

[23] Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 1, 2

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

[25] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019. 3

[26] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. Multi-scale image harmonization.

ACM Transactions on Graphics (TOG), 29(4):1–10, 2010. 1, 2

[27] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 5

[28] Michael W Tao, Micah K Johnson, and Sylvain Paris. Error-tolerant image compositing. In *European Conference on Computer Vision*, pages 31–44. Springer, 2010. 1, 2

[29] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3789–3797, 2017. 2, 3

[30] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 5

[31] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. *arXiv preprint arXiv:2207.04788*, 2022. 2, 3

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

[33] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 2022. 3

[34] Hui Zeng, Jianrui Cai, Lida Li, Zisheng Cao, and Lei Zhang. Learning image-adaptive 3d lookup tables for high performance photo enhancement in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 3

[35] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5, 7

[36] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2019. 2, 5

[37] Jun-Yan Zhu, Philipp Krahenbuhl, Eli Shechtman, and Alexei A Efros. Learning a discriminative model for the perception of realism in composite images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3943–3951, 2015. 2