DOI: 10.1111/cgf.14921

COMPUTER GRAPHICS *forum*
Volume 42 (2023), number 6, e14921

Harmonized Portrait-Background Image Composition

Yijiang Wang,[1] Yuqi Li,[1,2] Chong Wang[1] and Xulun Ye[1]

[1]Ningbo University, Ningbo, China
{2011082322, liyuqi1, wangchong, yexulun}@nbu.edu.cn
[2]Zhejiang Engineering Research Center of Advanced Mass Spectrometry and Clinical Application, Ningbo, China

Abstract

Portrait-background image composition is a widely used operation in selfie editing, video meeting, and other portrait applications. To guarantee the realism of the composited images, the appearance of the foreground portraits needs to be adjusted to fit the new background images. Existing image harmonization approaches are proposed to handle general foreground objects, thus lack the special ability to adjust portrait foregrounds. In this paper, we present a novel end-to-end network architecture to learn both the content features and style features for portrait-background composition. The method adjusts the appearance of portraits to make them compatible with backgrounds, while the generation of the composited images satisfies the prior of a style-based generator. We also propose a pipeline to generate high-quality and high-variety synthesized image datasets for training and evaluation. The proposed method outperforms other state-of-the-art methods both on the synthesized dataset and the real composited images and shows robust performance in video applications.

Keywords: image composition, portrait harmonization, stylegan

CCS Concepts: • Computing methodologies → Image segmentation; Image processing; Mixed/augmented reality

1. Introduction

Image composition is a visual task that aims to overlay a foreground object on another background image. One of the most popular and ubiquitous classes of foreground objects is the human portrait. Preserving a selfie portrait and replacing the original background to generate a composited image can help users hide their privacy as the background images usually contain environmental information. Users can also apply portrait-background image composition for entertainment purposes due to the creative experience brought by the technique. In the last decade, portrait-background image composition finds broad applications in numerous scenes, ranging from advertising design and movie creation to video telephony and photo/video synthesis, and augmented reality (AR) [WLZ*18, QLH21, dSCMZ20]. In particular, virtual video meeting has become popular during the COVID-19 pandemic, and portrait-background composition has become an essential function to change the user's virtual meeting environment. All these applications demand a high degree of visual realism in the composite images.

Many factors can affect the realism of the composited images. Usually, the unrealistic senses come from either the apparent inconsistency or the geometric inconsistency between the portrait images and the background images. The different capturing conditions and devices of the two images are the major reason causing the inconsistency. To eliminate the visual inconsistency, plenty of deep-learning-based methods have been proposed to harmonize the composited images [GGZ*21, JZZ*21, LXS*21, CNZ*21, CZN*20, SPK21, HIF20, CP20, TSL*17]. The image harmonization methods adjust foreground objects to make them compatible with backgrounds. However, most of the methods are designed for general foreground objects and rarely consider the intrinsic feature of portraits, which results in poor performance for portrait-background image composition. Human portraits, as the most widely studied image content, contain abundant priors to support challenging generation tasks such as portrait relighting [ZHSJ19], portrait synthesis [QLH21, JLG*20], and expression mapping [SLT*19, RCV*19], and so forth.

In this paper, we present a deep network named HPBNet that **h**armonizes the composited images of **p**ortrait and **b**ackground with high realism in various cases (see Figure 1). Motivated by the success of Generative Adversarial Network (GAN) in portrait generation and the style transfer architecture in attribution editing, we first improve harmonization quality by embedding composited images into the latent space of StyleGAN, and show that the latent code is capable of controlling the appearance of portraits to adapt to various background images. Our method considers both contents and styles

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

delivered by EUROGRAPHICS DIGITAL LIBRARY
www.eg.org diglib.eg.org

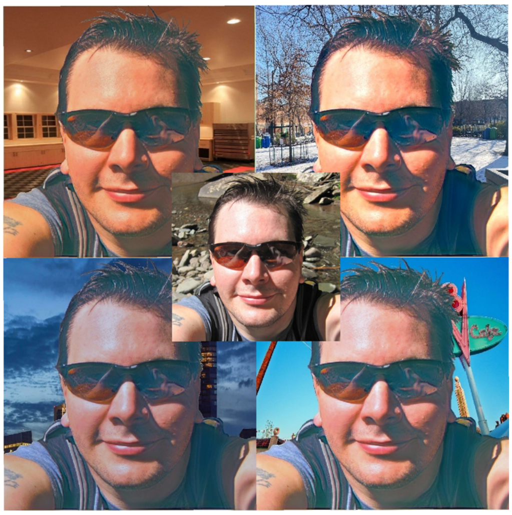

Figure 1: *Our portrait image harmonization method can enhance the realism of composited images by adjusting the foreground image to fit different backgrounds. The method not only harmonizes the colours of portraits and backgrounds but also reduces the inconsistency of the illumination. Note that the highlight on the portrait is relieved under the indoor scene (left upper).*

of the images and designs an end-to-end architecture to handle inconsistent colours and illuminations between foreground portraits and background images. We propose to apply the white balance to augment our dataset, and adopt colour transfer and portrait relighting techniques to increase the diversity of data for training. We validate the method on a synthesized dataset by comparison with the state-of-the-art image harmonization methods. We will release our code and pretrained models upon acceptance.

Our main contributions are summarized as follows:

- We *first* introduce a style-based generator to guide image harmonization. Our portrait-background image harmonization method significantly outperforms existing state-of-the-art methods in a variety of metrics.
- We present a portrait-background image synthesizing method to increase the diversity of the training set. Training on the synthesized dataset allows our network to harmonize the composited image with inconsistent illumination and colours.
- Experiments on the synthesized dataset and real images demonstrate the effectiveness and robustness of our image harmonization method.

2. Related Work

2.1. Image harmonization

Traditional methods mostly focus on prior knowledge of the background geometry and appearance. For example, Remez et al. [RHB18] find the reasonable position of foreground objects along the horizontal scan line and generate composited images without changing the scale of the foreground objects. Song et al. [SZQT20] propose an efficient method that models the foreground appearance transformation as channel-wise scales, and estimate the scales based on gray pixels of the source and the target background images to generate a harmonized image.

Deep-learning-based methods can extract higher-level features to harmonize the composited images. Tsai et al. [TSL*17] propose an end-to-end convolutional neural network for image harmonization, which can capture both the context and semantic information of the composite images during harmonization. Additionally, Cong et al. [CZN*20] present a domain verification discriminator, with the insight that the foreground needed to be translated to the same domain as the background. Convolutional neural networks can boost their representation capability for image harmonization with the success witnessed by attention modules, which are added to U-Net by Cun et al. [CP20] and Hao et al. [HIF20] to improve the harmonization effects. Guo et al. [GGZ*21] apply the self-attention mechanism and propose to adopt a transformer architecture for image harmonization. Ling et al. [LXS*21] treat image harmonization as a style transfer problem and propose a region-aware adaptive instance normalization module to apply the explicit visual style of the background to the foreground. Jiang et al. [JZZ*21] present a self-supervised harmonization framework to handle the problem without human-annotated masks or professionally create images for training, Sofiiuk et al. [SPK21] utilize the space of high-level features learned by a pretrained classification network to enhance the harmony of the composited images. More recently, a method [XRC*22] introduces a cutting-edge deep comprehensible colour filter (DCCF) algorithm, which utilizes four human comprehensible neural filters in conjunction with a deep learning model. Another study [XLW*22] learns the composition process from an imperfect alpha matte and adjusts colour distribution using appearance features extracted from the background. Additionally, a study [LCPW21] proposes a novel network for rendering spatially separated curves and incorporates the use of the cascaded and semantic modules for cascade refinement and semantic guidance, leading to a significant reduction in network parameters and improved results.

2.2. Portrait background replacement

To composite portraits into new scenes and enhance the realism of the composited images, portrait relighting techniques generate the composited images as though they are illuminated under new panoramic lighting environments. Sun et al. [SBT*19] capture a light stage dataset and propose a deep neural network to predict illumination and generate the relit portrait images. Zhou et al. [ZHSJ19] synthesize a portrait relighting dataset using a physically-based strategy, and improve the relighting quality by adding a GAN loss. Pandey et al. [PEL*21] design a complete system consisting of foreground estimation, relighting, and compositing to achieve automated portrait relighting and background replacement. Zhang et al. [ZZW*21] extend the task to the video domain, and propose an efficient video relighting method that eliminates flickering artifacts. However, only 6% of the panoramic scene is observed by a typical mobile phone camera, acquiring the panoramic lighting map from a single image is impractical [LMF*19]. Fortunately, the illumination of portraits is mainly from the direction of the camera, portrait images are rarely affected by the background lighting. Therefore, portraits-background image composition has a high tolerance for physical inaccuracy, and only demands visual consistency between portraits and background images.

2.3. Deep generative prior

Portrait-background harmonization is a special task of image-to-image translation, which takes images from one domain and transforms them to another domain to transfer image styles. Generative adversarial networks (GAN) [GPAM*14] consisting of a generative network and a discriminative network an effective tool to handle image-to-image translation tasks. The generative network learns the generation from the random samples in the latent space to the desired images. Especially in the field of human face generation, GAN has gained great research attention. More recently, Karras et al. [KLA19, KLA*20, KAL*21] develop a style-based generator architecture for GAN, named StyleGAN, which is inspired by the thought from style transfer literature, and the generator architecture can implicitly learn hierarchical latent codes in $\mathcal{W}$ latent space that contributed to the generated images. Kafri et al. [KPACO22] propose a latent code fusion method that supports the fusion of multiple attributes of portraits. Previous research and numerous experimental studies demonstrate that images generated through latent code control exhibit a diverse range of portrait properties across multiple dimensions. Compared with directly editing pixels, editing in the latent manifolds can keep the image located on the manifold and looks natural and realistic. This opens up the possibility to edit the attributes of the images while retaining the realism of the image. Applying StyleGAN to the portrait synthesis task, with latent code decoupling technique, either the coarse portrait attributes (e.g. pose, face shape, and so on) or the detailed attributes (e.g. pupil colour, hair colour, and so on) of portrait images can be separately controlled. Benefits from the characteristics of the method, high-realism and fidelity portrait editing becomes achievable. Nitzan et al. [NBLCO20] use a feature pyramid network to project input images into the $\mathcal{W}$ space, to generate highly identity images. Richard et al. [RAP*21] train multiple encoders to maintain both the features and identities of input images. Abdal et al. [AQW20] optimize the latent vector to minimize the error for the given image. With these achievements, deep generative priors can solve problems such as image restoration and deblurring on real images [WLZS21, HXZC21].

The goal of our portrait-background image harmonization problem is the same as for portrait editing. The appearances of portraits need to be adjusted according to the content of the backgrounds while the composited images should not fall off the manifold of the latent codes of the realistic portrait-background images. Therefore, we propose to apply StyleGAN to learn the hierarchical styles of real portrait-background images and utilize the multi-level style layer to guide the image harmonization. The detailed method is introduced in the next section.

3. Method

In this section, we will describe how to generate the synthesis dataset for training and present the architecture of our HPBNet framework. Given portrait image masks and the composited images that the appearances of the portrait and background are incompatible, our network is trained to generate harmonized images that ought to be close to ground truth. Our network consists of a U-net module and a StyleGAN generator module, which can both ensure data fidelity and force the composited images to satisfy style generative prior.

3.1. Dataset synthesis

Acquiring paired unharmonized and harmonized portrait-background images are infeasible. To construct a dataset for the training of our supervised-learning network, we generate the unharmonized images by changing the appearances of the portraits in real images and using the original real images as ground truth. Unlike previous image harmonization methods, we attempt to learn how to handle colour and illumination inconsistency. Therefore, the synthesized dataset must contain unharmonized images with wide colour and illumination variety as shown in Figure 2.

The original portrait dataset we use is FlickrFaces-HQ (FFHQ) [KLA19], which consists of 70,000 high-quality portrait-background images with 1024x1024 resolution, and contains a considerable variety in terms of age, ethnicity, and image background with a complete human face. In this work, we select 20,000 representative images from the FFHQ dataset. Note that the white balance of the images in FFHQ is similar, to avoid bias of white balance, we randomly adjust the white balance to introduce more colour variety to the source images in the dataset. The white balance is augmented by multiplying the images with random 3×3 diagonal matrices with values ranging from 0.75 to 1.3. Note that before and after the white balance augmentation, we apply a Gamma and inverse Gamma mapping to process the images with a Gamma value 2.2.

To create incompatible portraits, the first step is to segment the portraits from the source images to obtain the portrait masks. We utilize the portrait segmentation method proposed by [ZSQ*17] to generate the segmentation masks. Let us denote the source images as I_s, denote the segmented human portraits as H_s, and denote the mask as M_s.

Colour transfer. We begin by selecting a source image I_s and a random reference image I_r''. Our objective is to generate colour-transferred portraits P'. Instead of transferring colours in RGB space, we transfer colours in the nonlinear perceptual-metrics-based Lab space. After transferring H_i and H_j into Lab space, we subtract the mean Lab values on their corresponding channels to obtain the colour shift of each pixel in H_s and H_r, and the new colour shifts of the pixel in P' are obtained by multiplying the original colour shifts by the ratios between the standard deviations of the colour shifts of H_s and H_r as described in [RAGS01]. Finally, we convert P' from the Lab values to RGB space to obtain the colour-transferred portraits.

Portrait relighting. To increase the illumination diversity of the training data, we further relit the image P' with a random illumination to generate the image P for training. We randomly generate a sphere light-source image and feed it to the portrait relighting network proposed by [ZHSJ19] to apply the random lighting to the image P'. Note that only the portrait regions are relit, and we need to paste the processed portraits to the original background. More than 20,000 training data are generated to be fed into our network

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

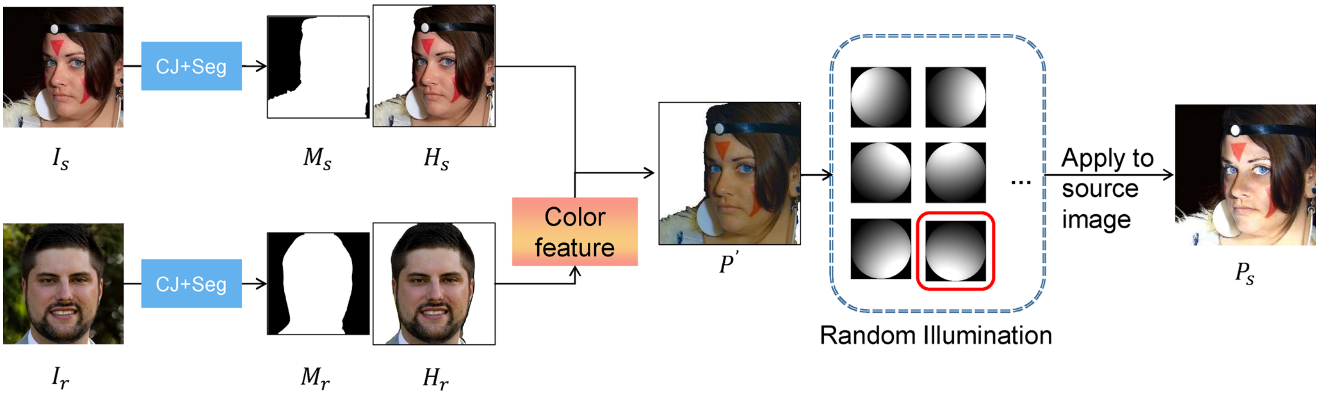

Figure 2: *Dataset synthesis. To generate the composited images that the appearances (colours, illuminations) of the portrait and the background are incompatible, we transfer the colours of the portraits in an original image by referring to the colours of another portrait, and then relight the intermediate image with with randomly oriented illumination.* ***CJ*** *denotes colour jitter processing that augments original images with white balance adjustment.*

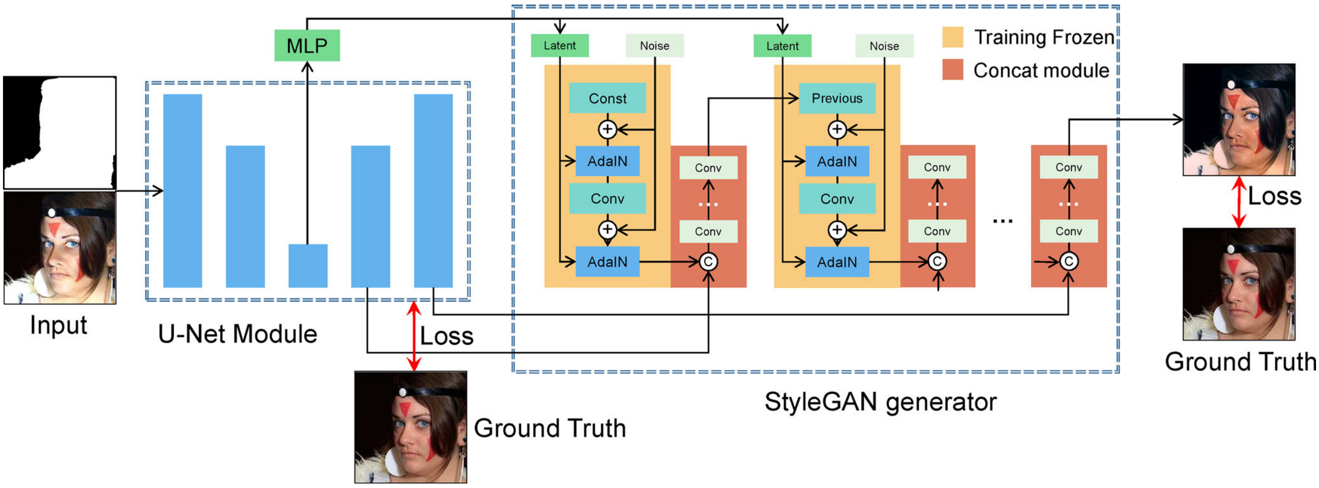

Figure 3: *The network architecture of HPBNet. The Network consists of a U-Net module and a StyleGAN generator module. Note that the extracted features of the U-Net bottleneck are decoupled and fed into the StyleGAN as inputs, and the feature maps of the multi-scale decoder of the U-Net are concatenated with the multi-scale layers of StyleGAN.*

to learn portrait-background image harmonization from the synthesized dataset.

3.2. U-Net module

As shown in Figure 3, the U-Net [RFB15] module of our network consists of multi-layer encoders and decodes with skip connections. The shallow layers learn low-level local features, while the deep layers learn high-level global features. The input of the U-Net is the paired composited image P and composition mask M with the same spatial resolution.

We adopt a structure across multiple scales in a ’coarse-to-fine’ strategy, that is, in the i-th scale of the U-Net, we take an initial harmonized result $\hat{I}^{i-1}$ (upsampled from the previous scale), and estimate the harmonized image $\hat{I}^{i}$ at this scale. Therefore, for each scale, we form the generation of a harmonized latent image as a sub-problem of the image harmonization task, the parameters of each scale are learned to represent the exact residual of the scale. The loss function of the U-Net is designed as:

$$\mathcal{L}_{unet} = \sum_{i=1}^{n} \frac{w_i}{N_i} ||\hat{I}^i - I^i||_1 + w_p \mathcal{L}_{perc}, \tag{1}$$

where $||\cdot||_1$ represents the L1 norm, $\mathcal{L}_{perc}$ denotes the perceptual loss which is formulated as the L1 norm of the features difference using VGG-19 as backbone with weight $w_p = 0.1$, I^i denotes the i-th downsampled image of ground truth source image I in the

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

dataset, n denotes the scale amount of the U-Net, $\{w_i\}$ denotes the weights for each scale, and N_i denotes the number of elements in I^i for normalization. We empirically set $w_i = 1.0$. The loss function takes into account both data fidelity and perceptual fidelity. The U-Net consists of seven scales of encoders and decoders. On each scale, the encoder and decoder contain three convolutional layers with the kernel size 3×3. The network uses LeakyReLU as the activation function.

Our U-Net is also a multiplexing architecture that provides latent codes to the StyleGAN module. The output feature maps of the bottleneck are sent to StyleGAN as the initial random vector inputs. In addition, the outputs of each scale are sent to the same level of StyleGAN as intermediate inputs. The utilization of the output from each scale of the U-Net serves the purpose of preserving the features of the final output image, thus avoiding the loss of personal identity in the portrait, which is a potential issue resulting from the prior of the StyleGAN. In addition, the U-Net allows us to make color adjustments for the unharmonized input images to a certain extent.

3.3. StyleGAN generator module

The StyleGAN generator module aims to generate a harmonized image from the features extracted by the bottom module of the U-Net. The structure of the StyleGAN generator consists of two parts: a multi-layer perceptron (MLP) mapping network, which decouples the features into a latent space representing a harmonized image-manifold of portrait-background; and a synthesis generator network consisting of multiple blocks, each controlling the style of the image at different levels.

The StyleGAN generator supports latent codes from the $\mathcal{W}$ latent space as input which allows for control over various facial attributes such as illumination, age, and gender. As our input is an image, which does not have a corresponding latent code for any real images. Therefore, it is necessary to project our real images into the $\mathcal{W}$ space, which is referred to as GAN inversion. In our HBPNet, we employ the encoder part of the U-Net as an auxiliary and use a multi-layer perceptron mapping network to project the encoding vectors into the $\mathcal{W}$ latent space. In this way, we obtain controllable data within the synthetic generation network, enabling the control of multiple facial attributes.

In the synthesis generative network, blocks receive only latent code and noise input, then output an image of the corresponding resolution as one of the inputs of the next layer. In our network, the upsampling network of each scale of the U-Net is sent to the corresponding synthesis scale for data concatenation. To keep the shape consistency of the input and output data in the synthesis network, we use a convolutional layer to reduce the channel amount of the concatenated tensors.

Portrait harmonization is to make the foreground portraits and the backgrounds harmonized and visually more realistic, so we employ the StyleGAN2 discriminator of the generative adversarial network to calculate the adversarial loss. We formulate it using the logistic loss:

$$\mathcal{L}_{adv} = -w_a \, \mathbb{E}\big(softplus(D(\hat{I}))\big), \tag{2}$$

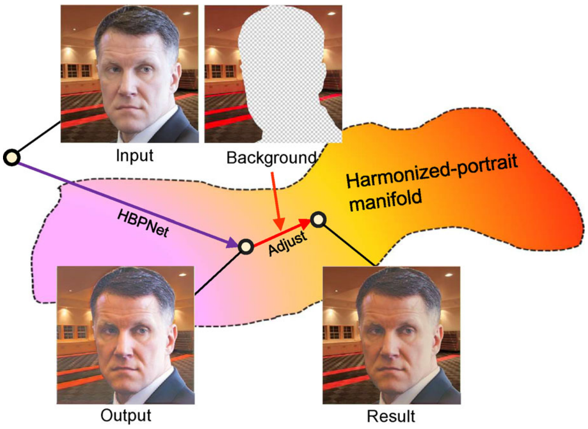

Figure 4: *The process of our portrait harmonization method. Our results satisfy both portrait image prior and colour constancy.*

where $softplus()$ denotes the smooth activation function for the discriminator $D()$, and w_a denotes the weight for the adversarial loss. We empirically set $w_a = 0.1$. Here the term $D(\hat{I}^i)$ is defined as the probability that the reconstructed image $\hat{I}^i$ is a realistic portrait image.

Although StyleGAN prior can enforce that the harmonized results fall on the manifold of real portrait images, our network cannot guarantee that the background of the direct-output results is precisely the same as that of the input. Instead, the network tends to jointly adjust the foreground portraits and background so that the output can fall on the harmonized portrait manifold. Therefore, we need a slight adjustment to match the colour temperature of the desired background and the direct-output background of the network and then apply the scale of the colour temperature adjustment to the output foreground to obtain the final result. Because our input training images contain various colour temperatures, there is a high probability that the final result will still fall on the manifold as shown in Figure 4. Here is the mathematical description of the colour temperature adjustment:

$$Result = \frac{\sum [I_s \cdot (1 - M_s)]}{\sum \big[\hat{I} \cdot (1 - M_s)\big]} \times \hat{I} \tag{3}$$

In summary, our network structure mainly consists of two parts, the U-Net module and the StyleGAN generator. The overall model objective is a combination of the above losses:

$$\mathcal{L} = \mathcal{L}_{adv} + \mathcal{L}_{unet} \tag{4}$$

4. Experiments

In this section, we present the implementation details of our experiment and conduct an ablation study to illustrate the progressiveness of our model. Subsequently, we compare our approach quantitatively and qualitatively with state-of-the-art techniques, and we supplement our findings with a user study to obtain subjective scoring. Additionally, we explore the applicability of our model to the video domain and perform an analysis of the results. Further experimental results are provided in the supplementary material.

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

4.1. Implementation details

In this section, we describe the datasets used during the experiments, the metrics measured, and the training details.

Datasets. We train our HBPNet using the FlickrFaces-HQ (FFHQ) dataset as the base dataset. We select 20,000 images for training and reserved the rest for inference. During the training process, for each selected source image I_s, we randomly choose another reference image I_r for colour transfer, followed by random lighting to generate a synthesis image P_s, as described in Section 3.1.

Metrics. To objectively evaluate our results, we compare the performance using two classic image quality assessment metrics, mean square error (MSE) and peak signal-to-noise ratio (PSNR). Additionally, we employ two deep metrics, learned perceptual image patch similarity (LPIPS) [ZIE*18] and deep image structure and texture similarity (DISTS) [DMWS20], to measure the results.

Training. We adopt the Pytorch framework to implement our models. We resize input images to 512 × 512 for dataset synthesis to reduce memory consumption. Training samples are augmented with random horizontal flip and white balance adjustments. To ensure the shape consistency of the layers, we use a seven-layer encoding-upsampling network structure for the U-Net, and also a seven-layer structure for the synthetic network of StyleGAN. For the MLP, we implement an eight-layer fully connected network. We use pre-trained weights for the StyleGAN trained on FFHQ which are fixed in the training process. Then we train the U-Net module and the concatenate layers. The model is trained for 60,000 iterations using Adam optimizer with a learning rate of 2×10^{-3}. Our training takes nearly 20 h on a single Tesla V100 GPU. The proposed method can harmonize a 512 × 512 image in 0.218 s.

4.2. Ablation studies

In this section, we investigate the contributions of the two components of our model and provide evidence to justify our final colour temperature conversion in real composited images.

Many existing deep learning-based image harmonization methods introduced U-Net as the backbone network [CZN*20, CNZ*21] due to its effectiveness in colour correction of unharmonized images. However, it might be impossible to correct the lighting effect on portrait images with a local convolutional neural network such as U-Net. To explore whether the U-Net network can be effective for unharmonized images where lighting and colour are both unharmonized, we construct a U-Net with a structure consistent with that used in HBPNet and use the same loss function and other settings. We compare the output results of a single U-Net and our network, and the detailed quantitative results can be found in Table 1.

It can be observed that the harmonization brought by U-Net is not significant, while the addition of StyleGAN leads to greater improvements in all metrics. Next, we visualize three examples of harmonized images in Figure 5. Due to the skip connections in U-Net, the identity features of the subjects and details of the background are well preserved, but the input colours and lighting are not processed correctly. [LFH21] draws a similar conclusion that pure CNN is not suitable to eliminate the non-uniform illumination effect. Our method combining U-Net and StyleGAN can provide promising results with harmonious colours and illuminations.

Table 1: *The improvement brought by U-Net module and StyleGAN generator module.*

Metrics	Unharmonized	U-Net only	U-Net + StyleGAN
MSE↓	2711.1	2398.76	**1181.86**
PSNR↑	14.50	14.88	**18.40**
LPIPS↓	0.21	0.24	**0.14**
DISTS↓	0.18	0.16	**0.13**

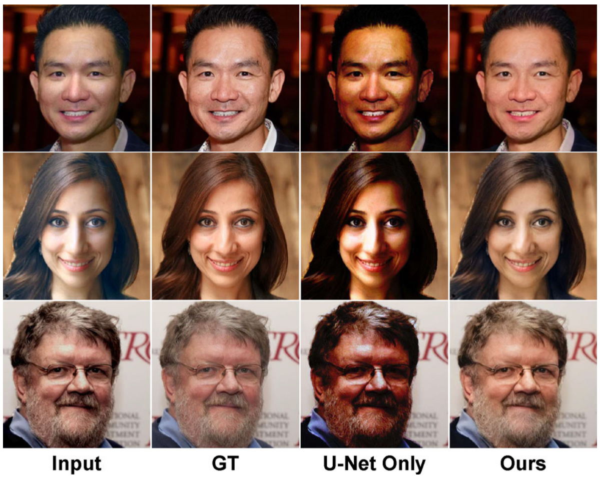

Figure 5: *The visual comparison of the unharmonized input images, ground truth, and harmonized results of U-Net and our network. Note that U-Net cannot handle the inconsistent lighting.*

Note that in the real composited image, our network can only guarantee that the results fail on a harmonized portrait manifold, which may lead to a minor change in the background. Therefore, we adjust the colour temperature in the final step.

4.3. Comparison with existing methods

Firstly, We compare our portrait-background image harmonization method with five state-of-the-art image harmonization methods on the synthesized images. The compared methods include ADFM [HIF20] , HT [GGZ*21], iDIH-HRNet [SPK21], S2CRNet [LCPW21], and DoveNet [CZN*20]. We not only utilize their provided pretrained models for inference but also retrain their models on the synthesis dataset generated as proposed in our work. The results of the six methods on the four metrics are shown in Table 2. The retrained models are better than the previous pretrained models in terms of metrics and visual results. Our method outperforms the other methods on all metrics by a large margin, although their retrained network show improvement to their pretrained network. Note that in the following sections, we use the retrained model of each method for comparisons.

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

Table 2: *Quantitative comparison of the harmonized results on our dataset in terms four metrics.*

Methods		MSE↓	PSNR↑	LPIPS↓	DISTS↓
Unharmonized		2711.10	14.50	0.21	0.18
HT	pretrained	2334.01	14.89	0.31	0.22
	retrained	1748.94	16.17	0.30	0.22
DoveNet	pretrained	2502.20	14.40	0.30	0.21
	retrained	1828.26	16.20	0.25	0.19
ADFM	pretrained	2291.10	17.31	0.30	0.22
	retrained	1853.00	16.20	0.24	0.18
iDIH-	pretrained	2443.50	14.90	0.22	0.18
HRNet	retrained	1675.70	16.60	0.25	0.19
S2CRNet	pretrained	2360.83	15.18	0.31	0.22
	retrained	2142.82	15.82	0.23	0.18
Ours		**1181.86**	**18.40**	**0.14**	**0.13**

The visual results of the six methods are shown in Figure 6. Our harmonized images are much more realistic and closer to the ground truth than the other four methods. In terms of image detail, our network also has better results than other methods in generating hairs and wrinkles in portraits. In addition, our network can eliminate the “red-eye effect” (caused by the camera capturing light reflected from the retina behind the person’s eyes when using the flash at night and in dim light) in some images. More results will be given in the supplemental materials.

Next, we compare the proposed method with the five image harmonization methods on real composited images. We randomly composite the portraits from the FFHQ dataset and 25 background images and manually choose 500 composited images with a low realism degree for the test. Since the harmonization for real composited images has no ground truth, we visualize seven examples of the results in Figure 7 to compare the performance of the six methods. The effectiveness of the compared methods in handling lighting and preserving portrait details is evaluated. Results show that while HT partially adjusts the brightness of the foreground portrait, it fails to protect portrait details and could not coordinate with some backgrounds. DoveNet and S2CRNet adjust the colours of the portrait regions but cannot remove the inharmony light artifacts. ADFM and iDIH-HRNet eliminate over-saturate in some regions, but still unable to handle the inconsistent light. Our method is the most realistic as it considers the impact of background color and preserves the natural appearance of the portraits.

4.4. Perceptual user study

In the previous subsection, we carried out comparative experiments with models in the synthesized images and real-composited images, so we try to keep parallel to the above experiments in the subjective

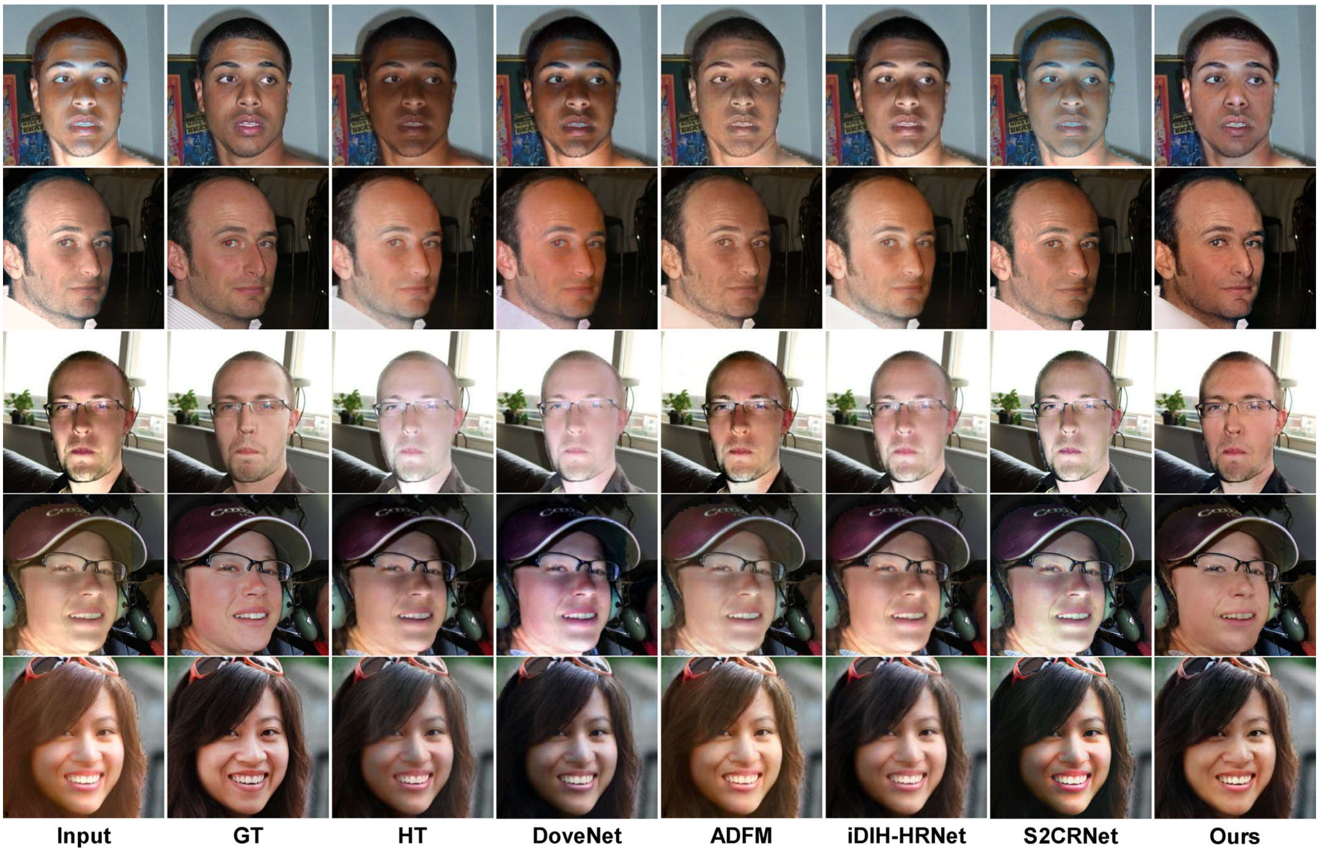

Figure 6: *Visual comparisons of the results of the six different methods on the synthesized images.*

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

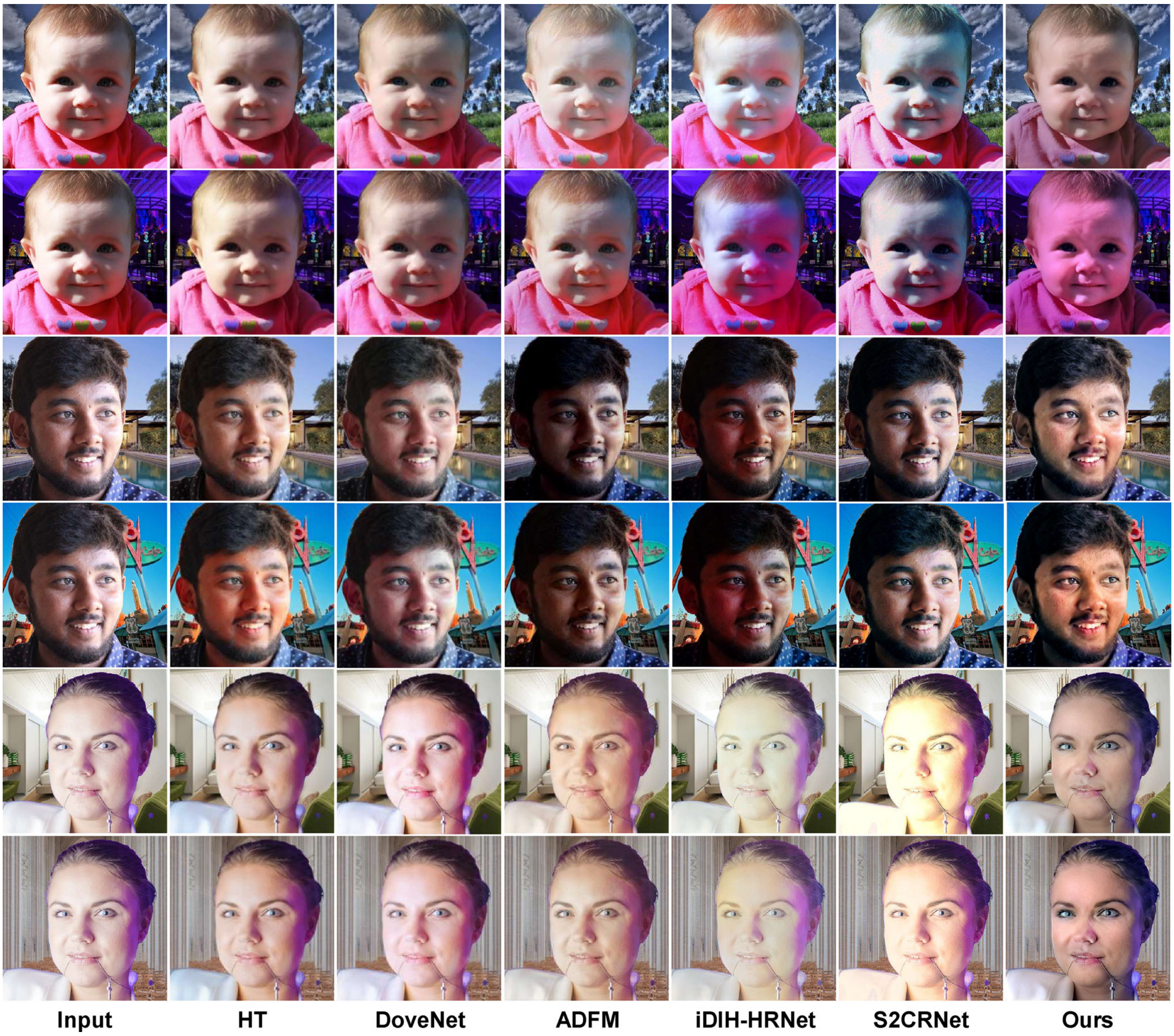

Figure 7: *Visual comparisons of the results of the six different methods on the real composited images.*

evaluations for comparison against five other methods. Nearly 200 sets of synthesized images, harmonized by the six methods, are selected to be evaluated. We invite 40 individual raters to evaluate the harmony and realism of each set. We also include 1375 sets of real composite images generated by different methods and ask raters to score them subjectively. To avoid bias, we randomly select two images from each set and presented them to the raters, who then select the one they believe to be more harmonious and realistic. we use the Bradley-Terry model (B-T model) [BT52, CZN*20] to calculate the global ranking scores for each method based on the final selection results.

The results are shown in Table 3. The ranking of our method is the closest to that of the ground truth, indicating that our results are as realistic as the ground truth. In the real composite images set, we also get the highest B-T score. We give more results of different methods in the supplementary document.

4.5. Extension to video

Image harmonization methods can be extended to composite dynamic scenes by directly processing each video frame as independent images. We argue that in a stable harmonized video, foreground portraits should exhibit changes in sync with the background, and the average brightness changes of the foreground and foreground should be consistent in the frequency distribution. A stable portrait image harmonization method should generate the video frames without noticeable flickers. To compare the stableness of different methods, we select multiple videos from the SumMe dataset

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

Table 3: *B-T scores of different methods on the synthesized images and real composite images.*

Method	B-T score↑	
Composite real images	–	1.94
Ground truth	<u>8.29</u>	–
HT	5.18	6.27
DoveNet	4.01	5.99
ADFM	5.89	4.83
iDIH-HRNet	5.83	4.25
S2CRNet	4.02	4.39
Ours	**7.78**	**6.33**

The first column of B-T score is the user study result of synthesized images, which has a ground truth in this experiment. We have underlined it as the actual best score. The second column indicates the results of real composite images.

as background videos and select multiple portraits from the FFHQ dataset as foreground, and composite videos frame-by-frame by using the six methods.

In order to quantify the stability of the generated videos, we try to perform a Fourier transform on the average brightness profiles of the foreground and background of each video and observe their energy distributions in the frequency domain. The energy distribution in the frequency domain of the average brightness of the foreground and background of a stable video should be similar. We present the results in Figure 8. The frequency distribution curves of our method and iDIH-HRNet are closest to those of the background. According to our hypothesis, this indicates that the foreground variation of these two methods is more consistent with the background variation than others. For the methods such as HT and DoveNet, whose curves have a very low magnitude of high frequency, indicating that they are not sensitive to the background changes. While for the methods such as ADFM and S2CRNet, the magnitude of the high frequency is much higher than that of the background, indicating that they may cause undesired flickers. Our method can be directly applied to video applications frame-by-frame, which currently achieve 5 fps. The composited videos will be provided in the supplemental materials.

5. Discussion and Conclusion

5.1. Limitations

Our method remains three major technical limitations. First, the method cannot handle extremely saturated illumination since our augmented dataset does not contain such cases. This would be improved if the dataset could be augmented with more extreme white balance scales. Besides, our method currently can only process portrait foregrounds. We will seek harmonization methods for human-body images, as well as image harmonization for other objects in the future. Finally, the efficiency of our method is not enough for a real-time application. We need to find lighter architectures like MobileStyleGAN [Bel21] to enhance efficiency.

5.2. Conclusion

We present a novel end-to-end portrait-background image harmonization network for enhancing the degree of realism of the composited images. Our method is capable to adjust the foreground portraits to fit the appearance of the background image while preserving the normal appearance of the portraits. Unlike previous methods, the proposed method can handle the inconsistency of both colours and illuminations. We present a dataset synthesis method to generate a large-scale, high-quality image dataset for training and evaluation. The experiments show that our method can provide more realistic, robust results than other state-of-the-art methods.

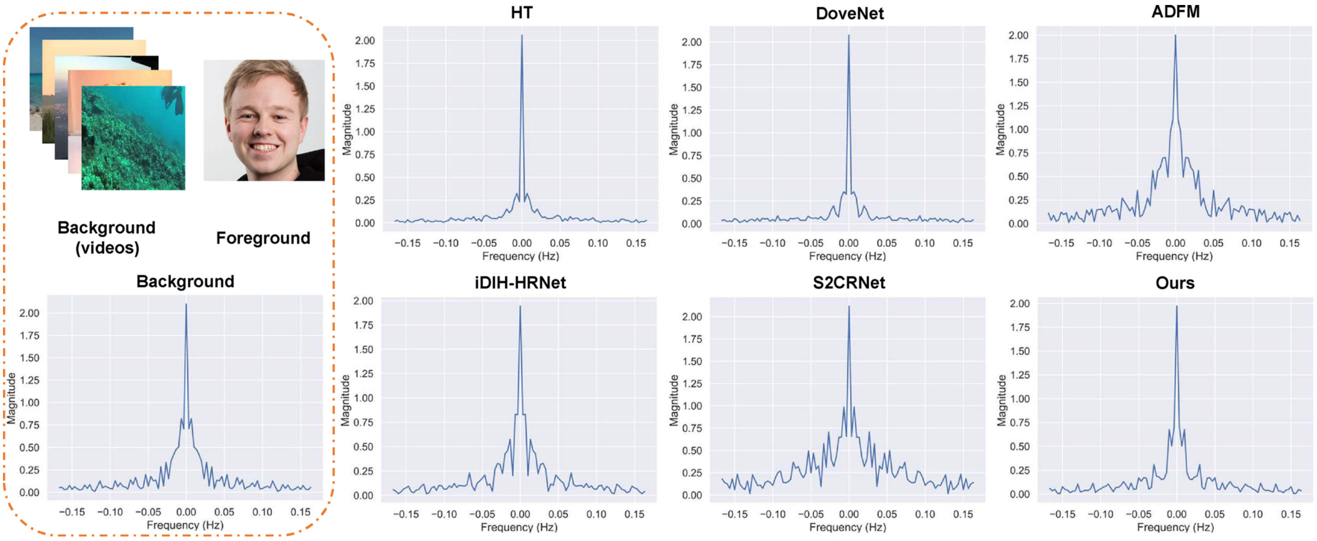

Figure 8: *The comparisons of the frequency distribution of the background and the harmonized foreground of the six methods. The curves of ours and iDIH-HRNet are the first two closest to the curve of the background.*

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

Acknowledgements

The work has been supported by the Ningbo Innovation Challenge Project (No. 2022T001) and the Natural Science Foundation of Zhejiang Province (Grant No. Y19F020050).

References

[AQW20] Abdal R., Qin Y., Wonka P.: Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8296–8305.

[Bel21] Belousov S.: Mobilestylegan: A lightweight convolutional neural network for high-fidelity image synthesis, 2021. arXiv:2104.04767.

[BT52] Bradley R. A., Terry M. E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika 39, 3/4* (1952), 324–345.

[CNZ*21] Cong W., Niu L., Zhang J., Liang J., Zhang L.: Bargainnet: Background-guided domain translation for image harmonization. In *2021 IEEE International Conference on Multimedia and Expo (ICME)* (2021), IEEE, pp. 1–6.

[CP20] Cun X., Pun C.-M.: Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing 29* (2020), 4759–4771.

[CZN*20] Cong W., Zhang J., Niu L., Liu L., Ling Z., Li W., Zhang L.: Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8394–8403.

[DMWS20] Ding K., Ma K., Wang S., Simoncelli E. P.: Image quality assessment: Unifying structure and texture similarity. *CoRR abs/2004.07728* (2020). URL: https://arxiv.org/abs/2004.07728.

[dSCMZ20] de Souza Cardoso L. F., Mariano F. C. M. Q., Zorzal E. R.: A survey of industrial augmented reality. *Computers & Industrial Engineering 139* (2020), 106159.

[GGZ*21] Guo Z., Guo D., Zheng H., Gu Z., Zheng B., Dong J.: Image harmonization with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 14870–14879.

[GPAM*14] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.: Generative adversarial nets. *Advances in Neural Information Processing Systems 27* (2014).

[HIF20] Hao G., Iizuka S., Fukui K.: Image harmonization with attention-based deep feature modulation. In *The British Machine Vision Conference (BMCV)* (2020).

[HXZC21] He Y., Xing Y., Zhang T., Chen Q.: Unsupervised portrait shadow removal via generative priors. In *Proceedings of the 29th ACM International Conference on Multimedia* (2021), pp. 236–244.

[JLG*20] Jiang W., Liu S., Gao C., Cao J., He R., Feng J., Yan S.: Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020).

[JZZ*21] Jiang Y., Zhang H., Zhang J., Wang Y., Lin Z., Sunkavalli K., Chen S., Amirghodsi S., Kong S., Wang Z.: Ssh: A self-supervised framework for image harmonization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 4832–4841.

[KAL*21] Karras T., Aittala M., Laine S., Härkönen E., Hellsten J., Lehtinen J., Aila T.: Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems 34* (2021), pp. 852–863.

[KLA19] Karras T., Laine S., Aila T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 4401–4410.

[KLA*20] Karras T., Laine S., Aittala M., Hellsten J., Lehtinen J., Aila T.: Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8110–8119.

[KPACO22] Kafri O., Patashnik O., Alaluf Y., Cohen-Or D.: Stylefusion: Disentangling spatial segments in stylegan-generated images. *ACM Transactions on Graphics (TOG)* (2022).

[LCPW21] Liang J., Cun X., Pun C.-M., Wang J.: Spatial-separated curve rendering network for efficient and high-resolution image harmonization, 2021. arXiv:2109.05750.

[LFH21] Li Y., Fu Q., Heidrich W.: Multispectral illumination estimation using deep unrolling network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 2672–2681.

[LMF*19] LeGendre C., Ma W.-C., Fyffe G., Flynn J., Charbonnel L., Busch J., Debevec P.: Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5918–5928.

[LXS*21] Ling J., Xue H., Song L., Xie R., Gu X.: Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 9361–9370.

[NBLCO20] Nitzan Y., Bermano A., Li Y., Cohen-Or D.: Face identity disentanglement via latent space mapping. *ACM Transactions on Graphics (TOG) 39* (2020), 1–14.

[PEL*21] Pandey R., Escolano S. O., Legendre C., Häne C., Bouaziz S., Rhemann C., Debevec P., Fanello S.: Total

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

relighting: Learning to relight portraits for background replacement. *ACM Transactions on Graphics 40*, 4 (Jul 2021).

[QLH21] Qi M., Li Y., Heidrich W.: Isp-agnostic image reconstruction for under-display cameras. *arXiv preprint arXiv:2111.01511* (2021).

[RAGS01] Reinhard E., Adhikhmin M., Gooch B., Shirley P.: Color transfer between images. *IEEE Computer Graphics and Applications 21*, 5 (2001), 34–41.

[RAP*21] Richardson E., Alaluf Y., Patashnik O., Nitzan Y., Azar Y., Shapiro S., Cohen-Or D.: Encoding in style: A stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021).

[RCV*19] Rössler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Nießner M.: FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)* (2019).

[RFB15] Ronneberger O., Fischer P., Brox T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention* (2015), Springer, pp. 234–241.

[RHB18] Remez T., Huang J., Brown M.: Learning to segment via cut-and-paste. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 37–52.

[SBT*19] Sun T., Barron J. T., Tsai Y.-T., Xu Z., Yu X., Fyffe G., Rhemann C., Busch J., Debevec P. E., Ramamoorthi R.: Single image portrait relighting. *ACM Transactions on Graphics 38*, 4 (2019), 79–1.

[SLT*19] Siarohin A., Lathuilière S., Tulyakov S., Ricci E., Sebe N.: First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)* (December 2019).

[SPK21] Sofiiuk K., Popenova P., Konushin A.: Foreground-aware semantic representations for image harmonization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2021), pp. 1620–1629.

[SZQT20] Song S., Zhong F., Qin X., Tu C.: Illumination harmonization with gray mean scale. In *Computer Graphics International Conference* (2020), Springer, pp. 193–205.

[TSL*17] Tsai Y.-H., Shen X., Lin Z., Sunkavalli K., Lu X., Yang M.-H.: Deep image harmonization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 3789–3797.

[WLZ*18] Wang T.-C., Liu M.-Y., Zhu J.-Y., Liu G., Tao A., Kautz J., Catanzaro B.: Video-to-video synthesis. *arXiv preprint arXiv:1808.06601* (2018).

[WLZS21] Wang X., Li Y., Zhang H., Shan Y.: Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 9168–9178.

[XLW*22] Xing Y., Li Y., Wang X., Zhu Y., Chen Q.: Composite photograph harmonization with complete background cues. In *Proceedings of the 30th ACM International Conference on Multimedia* (2022), MM '22, Association for Computing Machinery, New York, NY, USA, pp. 2296–2304. URL: https://doi.org/10.1145/3503161.3548031.

[XRC*22] Xue B., Ran S., Chen Q., Jia R., Zhao B., Zhao B.: Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *ECCV* (2022).

[ZHSJ19] Zhou H., Hadap S., Sunkavalli K., Jacobs D. W.: Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 7194–7202.

[ZIE*18] Zhang R., Isola P., Efros A. A., Shechtman E., Wang O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (2018).

[ZSQ*17] Zhao H., Shi J., Qi X., Wang X., Jia J.: Pyramid scene parsing network. In *CVPR* (2017).

[ZZW*21] Zhang L., Zhang Q., Wu M., Yu J., Xu L.: Neural video portrait relighting in real-time via consistency modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 802–812.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information

© 2023 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.