Composite Photograph Harmonization with Complete Background Cues

Yazhou Xing
The Hong Kong University of Science and Technology

Yu Li*
International Digital Economy Academy

Xintao Wang
Applied Research Center, Tencent PCG

Ye Zhu
Applied Research Center, Tencent PCG

Qifeng Chen*
The Hong Kong University of Science and Technology

Figure 1: Our portrait composite results. We are the first to unify the objectives of both seamless blending and color harmonization in one pipeline. Compared with direct copy-and-paste method, our approach can produce both seamless boundaries and harmonic color in the composites.

ABSTRACT

Compositing portrait photographs or videos to novel backgrounds is an important application in computational photography. Seamless blending along boundaries and globally harmonic colors are two desired properties of the photo-realistic composition of foregrounds and new backgrounds. Existing works are dedicated to either foreground alpha matte generation or after-blending harmonization, leading to sub-optimal background replacement when putting foregrounds and backgrounds together. In this work, we unify the two objectives in a single framework to obtain realistic portrait image composites. Specifically, we investigate the usage of a target background and find that a complete background plays a vital role in both seamlessly blending and harmonization. We develop a network to learn the composition process given an imperfect alpha matte with appearance features extracted from the complete background to adjust color distribution. Our dedicated usage of a complete background enables realistic portrait image composition and also temporally stable results on videos. Extensive quantitative and qualitative experiments on both synthetic and real-world data demonstrate that our method achieves state-of-the-art performance.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision tasks**.

KEYWORDS

Image harmonization, Image matting, Deep learning

ACM Reference Format:
Yazhou Xing, Yu Li, Xintao Wang, Ye Zhu, and Qifeng Chen. 2022. Composite Photograph Harmonization with Complete Background Cues. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3548031

*Corresponding authors.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548031

1 INTRODUCTION

Portrait compositing aims at generating realistic photographs or videos by compositing portrait foregrounds and virtual backgrounds, which is a key technology in the image and video editing, film production, and video conferencing. Although artists can use photo editing software such as Adobe Photoshop to composite photos, it requires professional skills and is time-consuming to finish the task. Therefore an automatic and effective portrait compositing method is still in great demand. To make the composites realistic, a desirable method should precisely cut out the portrait foreground with refined edges and adjust the appearance of the foreground to better match the background. Previous research works in this direction separately dedicate to the two sub-tasks as 1) image matting to generate better foreground alpha matte and foreground color [2, 3, 7, 10, 11, 13, 15, 17, 19, 23–25, 27, 29, 31, 33, 35] to extract the foreground object layer, and 2) image harmonization to match the appearance differences (color, brightness, contrast, saturation etc.) of the foreground to the background [4, 6, 9, 18, 20, 28, 30, 32] caused by the lighting conditions, camera settings, and post-editings. However, we often get unrealistic composite results in such a two-step pipeline, as shown in Figure 3. We have investigated this and identified the following issues of this two-step approach.

(1) Although automatic matting (i.e., trimap-free) has made great progress recently using deep learning, clearly separating the foreground from the background is still challenging for alpha matting approaches, especially when the background is complicated and when there are transparent parts in the foreground such as human hairs and lace clothes. The inaccurate estimation in the matting and the color decontamination can result in artifacts along the foreground boundaries in the composites.
(2) Traditional harmonization methods [18, 20, 28, 32] rely on global statistics and have no awareness of the semantic content. They will sometimes generate unnatural colors in the results. Recently deep learning-based harmonization methods have been proposed and show better performance. They adjust the foreground appearance after compositing the foreground with the background. This setup can only use parts of the background information and lead to unsatisfactory and in-stable harmonization results.
(3) The lack of large-scale and high-quality real-world datasets also hinders image matting and harmonization performance in real-world cases.

To tackle this limitation, we propose a novel pipeline for compositing images and videos towards the realism of the final composites. Our method tries to obtain seamless blending around the foreground object boundary and properly adjust the appearance of the foreground in one framework. Our method is based on state-of-the-art matting techniques, and our input is a triplet of a source image or video, a target background, and an imperfect foreground mask extracted by an off-the-shelf alpha matting method. To alleviate the boundary artifact brought by inaccurate alpha estimation and linear alpha blending, we propose a learnable blending framework with deep neural networks. Besides, we show that a complete target background is vital to effectively extract the background appearance for image harmonization. We show that it also helps maintain the temporal consistency for video harmonization. We note that there is no existing large-scale dataset composed of real portraits and the accompanying backgrounds. Thus, to prove the effectiveness of our proposed compositing framework, we utilize one public group-level portrait retouching dataset [14] to synthesize the training data through compositing one source image onto another photo within the same group. We experimentally prove that our framework can generate better composite photos than state-of-the-art baselines that mostly separately handle the image compositing task. We show that our method can also obtain high-quality and temporally consistent results for video harmonization.

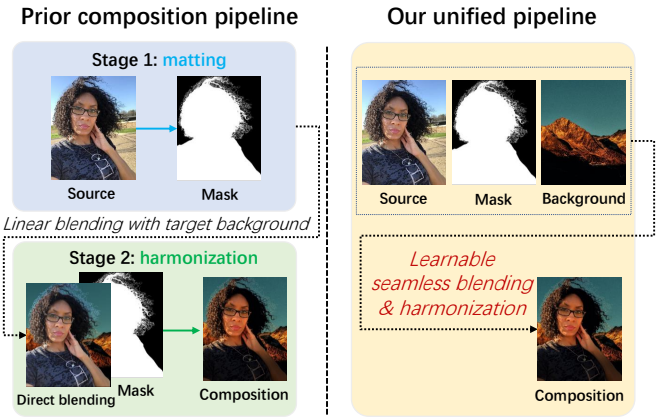

Figure 2: Comparison between our framework and previous methods. Prior work usually separates the composition task into alpha matting and after-blending harmonization. Linear blending is utilized to composite the foreground with the target background given the predicted alpha mask. We propose the unified pipeline with the learnable blending and harmonization process to provide both seamless blending and color harmonization.

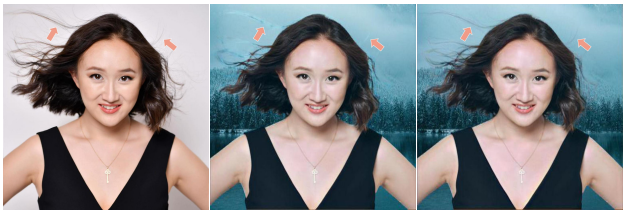

Figure 3: Compared with two-stage image composition (matting + harmonization), our unified pipeline can achieve more seamless blending results at boundaries and remedy detail loss of alpha mask. (Please zoom-in on the digital version)

2 RELATED WORK

2.1 Image Matting

Deep learning has shown great advantages in image matting problem [7, 10, 13, 15, 17, 19, 23, 25, 27, 29, 31, 33, 35]. Shen et al. [25] propose an automatic portrait matting method that is free of user interactions. DIM [31] is another early work that utilizes deep learning to natural image matting. To leverage both low-level features and high-level contexts, DIM takes both trimaps and images as input and defines a two-stage network to predict and refine the predicted alpha matte. Different from previous works, Hou et al. [7] propose to simultaneously predict foreground and alpha to obtain more accurate matting results, especially at edge regions. Due to the rapid growth of portrait photography, human matting has attracted great attention among image matting. Liu et al. [17] propose to leverage the coarse annotated mask to boost the performance of human matting. A similar idea has also been used in [33]. Recently, Ke et al. [10] propose MODNet that enables real-time trimap-free portrait matting. By taking MobileNetV2 as a backbone and the one-frame delay strategy, MODNet can produce temporal coherent video matting results in real-time. Zhang et al. [34] propose a joint pipeline to do image matting and image compositing. However, their method ignores the harmonization requirement of realistic image compositing. Ren et al. [21] propose to simultaneously optimize the image matting and harmonization with GAN. However, their method still relies on linear blending which will cause noticeable artifacts at boundary regions. Besides, their harmonization pipeline actually takes incomplete background as input, which will degrade the performance when the foreground occupies a large region at the final composites [9].

Thus, even though matting has achieved significant improvement over the years, the problem is still not completely solved yet and unpleasing artifacts usually appear at boundary regions. The inaccurate estimation of image matting can bring notable artifacts for further image composition. Besides, unifying the matting and harmonization towards final realistic composites is still challenging.

2.2 Image Harmonization

Image harmonization is a key technique for photo-realistic image editing. Traditional methods mainly focus on low-level statistics, such as color distribution [18, 20, 32] and multi-scale various statistics [28]. Deep learning techniques have greatly improved the quality of image harmonization [4–6, 9, 26, 30]. Tsai et al. [30] propose a deep neural network that can capture both semantic and context information to harmonize an image. Through training on the synthesized large-scale dataset, they outperform traditional methods. Cong et al. [4] contribute a public dataset iHarmony4 by generating synthesized composite images based on four public datasets. Furthermore, they propose an image harmonization method based on domain verification. Recently, researchers leverage intrinsic decomposition to solve image harmonization problem [6]. They propose a novel method through harmonizing reflectance and illumination intrinsic images, respectively. Jiang et al. [9] propose a self-supervised method to learn the separation of content and style of a pair of images. Their method can get rid of masks during training time.

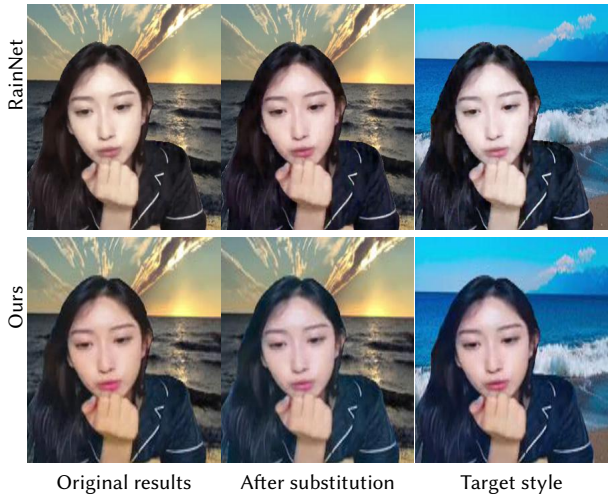

Figure 4: Analysis on the effectiveness of style control module in RainNet [16] and ours. We substitute the mean and variance parameters in RainNet to another styles' values. We find their results are inconsistent with the target style but ours are consistent.

However, existing image harmonization methods require a perfect foreground alpha layer when applied to an inharmonic image, which is not practical in reality.

3 PRIOR WORK ANALYSIS

We believe that complete target background is an important component for realistic image compositing, which is overlooked by recent works. In this section, we first revisit the typical problem setting of image harmonization in recent publications, and then we analyze the disadvantages of prior works due to the lack of complete target backgrounds.

As shown in Figure 2, prior works [4, 16] mainly address the problem through perturbing the foreground color of real photos and training deep networks to learn to restore the original foreground appearance (color, tone, illumination, etc.) in the photos. We argue that this kind of design may face some potential issues for realistic compositing.

Firstly, a realistic composite requires seamless blending, which cannot be achieved without complete backgrounds. Prior harmonization works require linear blending with perfect masks, which is hard to achieve in reality. As a consequence, linear blending with such an imperfect mask will produce noticeable artifacts on the boundaries, as shown in Figure 7. In this work, we show that we can utilize a learning-based blending method to replace the simple linear blending process, which can effectively alleviate the boundary artifacts, as shown in Figure 7.

Secondly, extracting the appearance from the partial background restricts the harmonization performance for the image and will produce inconsistent harmonization results when applied on videos. We notice that prior methods cannot produce temporally consistent results even when we compose a moving foreground onto a static background. We assume that this is due to the distorted incomplete background affecting the network to produce consistent results.

Figure 5: Illustration of our generated dataset. Given two photos shot at close time and the same scene, we composite the foreground of one photo to the other one and assume the composites are harmonic for the supervised training. We also composite the foreground of the source photo to novel backgrounds and utilize color transfer [20] to generate the input of our harmonization network, as shown in 'Input image'.

We take RainNet [16] to verify our assumption. RainNet explicitly formulates the harmonization problem as a style transfer task and utilizes masked adaptive instance normalization (AdaIN) [8] to transfer the appearance of background to foreground. Surprisingly, we find that if we substitute the mean and variance parameters of AdaIN in RainNet to another styles' values, the results are inconsistent with the target style. Examples are shown in Figure 4. This implies that the background appearance and foreground content are still entangled instead of being separated by their proposed region-aware AdaIN, which further hinders the network from producing temporally consistent results on videos.

Different from prior work, we show that with complete backgrounds, our model can produce seamlessly-blended and globally harmonic results on images and temporally consistent results on videos.

4 METHOD

We denote the input portrait image as I_s, the target background image as I_{bg}, the intermediate compositing image as I_c, the predicted alpha map as I_α, and the ground-truth harmonized composite image as I_h. Our network takes I_s, I_α, and I_{bg} as input, and learns to predict the compositing image $\hat{I}_c$ and ultimate harmonized compositing image $\hat{I}_h$. We will analyze the challenges of preparing such a dataset for training in Section 4.1.

4.1 Realistic data generation

We first list some training data requirements for our task.

1) I_h should be real photos instead of synthetic ones, which is vital to provide realistic supervision of the model. We also notice that existing public matting datasets only contain foreground and alpha masks and thus cannot be used for training our framework.
2) I_{bg} should be a complete image without any holes on it, i.e., I_{bg} cannot be the image with its foreground segmented out. Although previous image harmonization methods take such backgrounds as input, we claim that having the complete background is important to extract the appearance information for realistic compositing.

However, to meet both requirements 1) and requirement 2), we need a dataset that contains numerous pairs of the complete background and the person(s) before that background, which is hard to collect in reality. In this paper, we take a non-trivial attempt to construct a high-quality realistic dataset that can be a proper alternative to such a real-world dataset. Specifically, we utilize the expert-retouched portrait photo groups of [14] to construct our dataset, where each group is retouched with the same tone. Then we compose the foreground of one photo to another one in the same group. Due to the group-level consistency characteristics of the dataset, we can reasonably assume the composited photographs are harmonic. To generate the input photos, we additionally collect 1193 background images and apply color transfer to transfer the color information from backgrounds to the foregrounds, which is similar to [4, 30]. An illustration of our dataset construction process can be seen as Figure. 5. In practice, we use the public PPR10K dataset and randomly split one group in the dataset as foreground and background images, and we recompose the foreground people to the background ones.

4.2 Blending network

We notice that the boundary values of the predicted alpha mask and foreground color layer by existing alpha matting methods are hard to be perfectly accurate. Thus, simple linear alpha blending adopted in plenty of existing works [4, 9, 16] will introduce unpleasing artifacts such as color contamination artifact, as shown in Figure. 7. In this work, we present a CNN-based blending method that can produce composition results seamlessly with the inaccurate alpha masks predicted by existing pretrained matting networks, which can get rid of the color contamination issue at the boundary.

To train our blending network, we collect about 500 pairs of portrait foreground and alpha masks from the public matting dataset [12, 19, 31]. Following the data synthesizing pipeline in [10, 31], we also collect 140 diverse backgrounds and composite the foregrounds with alpha blending. In total, we construct 46K images for training and test. We randomly selected 45K images for training, 200 images for validation, and 900 images for the test set. We design our blending network as a shallow U-net [22], as shown in Figure. 6. To enable our blending network to handle the boundary blending

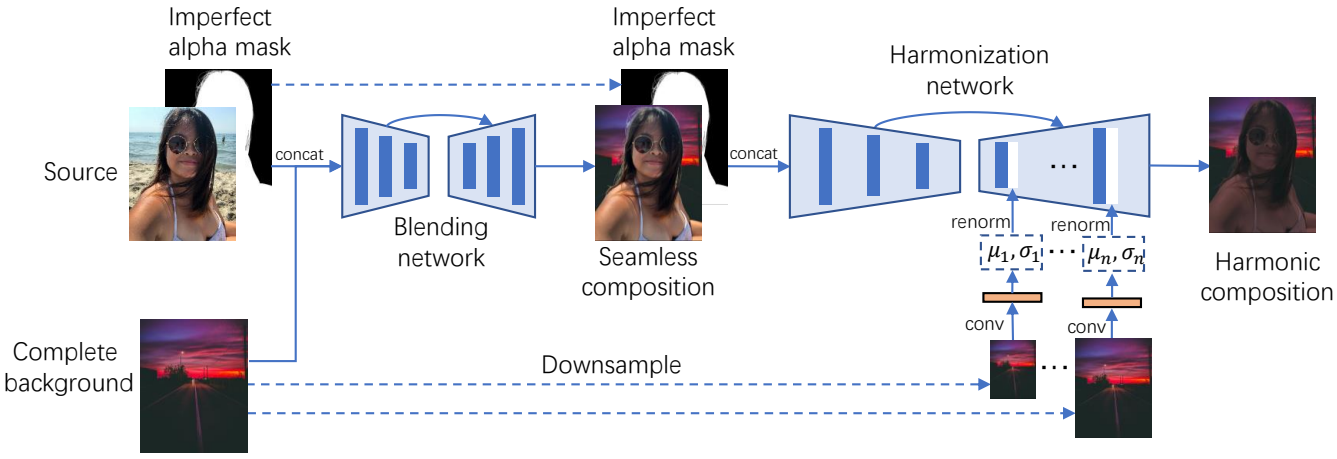

Figure 6: The overview of our framework. Our unified composition pipeline consists of a blending network and a harmonization network. Given a source photo with its imperfect alpha mask predicted by existing matting methods, our blending network learns to seamlessly blend the foreground and target background. Our harmonization network thus takes the seamless composites and the imperfect alpha mask as input and produces the harmonic results. Different from prior works, our framework can extract the appearance feature from the complete background instead of partially occluded ones.

better, we train the network with both composition loss as well as boundary loss [7, 10]

$$L_{blend} = ||\hat{I}_c - I_c||_1 + m_d||\hat{I}_c - I_c||_1, \tag{1}$$

where m_d is boundary area generated through erosion and dilation operations [10].

4.3 Harmonization network

Different from previous works [4, 16] that only extract the appearance feature from the masked partial background regions, our harmonization framework has some advantages with the complete target backgrounds. We claim two advantages of our new harmonization framework:

1) Being aware of the entire background helps the network to learn better appearance information from the background and thus produce more harmonic results than previous works [4, 16], especially when the foreground occupies a large part of the photo, which has also been identified in [9].
2) After training on an image harmonization dataset, our framework can be directly applied to video harmonization to produce temporal consistent harmonization results. We experimentally show that previous image harmonization methods often generate temporally flickering results when applied to videos.

We use U-net [22] as our backbone network. Similar to [4], we utilize three attention blocks in the U-net decoder. Unlike previous works, we feed the complete backgrounds to each layer of the decoder to provide the appearance feature to the foreground. Specifically, We develop two convolutional layers with ReLU activation functions [1] at each decoder scale to extract appearance features from the complete backgrounds. Then inspired by AdaIN [8], we feed the mean and variance parameters of each background feature to normalize the foreground area:

$$\bar{F}^i_{h,w,c} = \mu^i_c(s(I_{bg}))\frac{F^i_{h,w,c} - \mu^i_c}{\sigma^i_c} + \sigma^i_c(s(I_{bg})), \tag{2}$$

where μ^i_c and σ^i_c are the channel-wise mean and variance of the foreground feature $F^i_{h,w,c}$. $\mu^i_c(s(I_{bg}))$ and $\sigma^i_c(s(I_{bg}))$ are the mean and variance of the learned modulation feature map of I_{bg}. s is the feature extraction function implemented by two convolutional layers with ReLU activation.

We adopt reconstruction loss and adversarial loss to train our network:

$$L_{harmon} = L_{rec} + \lambda L_{adv}, \tag{3}$$

where $L_{rec} = ||\hat{I}_h - I_h||_1$ and L_{adv} is adopted from Cong et al. [4].

Discussion on the possibility of a one-stage framework. We design a two-stage network for image compositing due to data limitation. Our generated dataset in Sec. 4.1 has greatly alleviated the data shortage of image matting and harmonization. However, it still requires a large-scale dataset with plenty of triplets of (I_s, I_{bg}, I_h) for a one-stage framework, whose requirement is hardly to meet in reality. Thus, our two-stage framework is a more feasible solution and we demonstrate its good performance for image compositing with experiments.

5 EXPERIMENTS

5.1 Experimental setup

Datasets We use PPR10K [14] to construct our dataset for training and evaluation. There are 11,161 portrait photos in 1681 groups. We note that the raw photos within one group may vary a lot due to the difference in capturing settings and the environmental lighting conditions. Thus, we take the manual retouched results of PPR10K, whose objective is to preserve human-region priority and

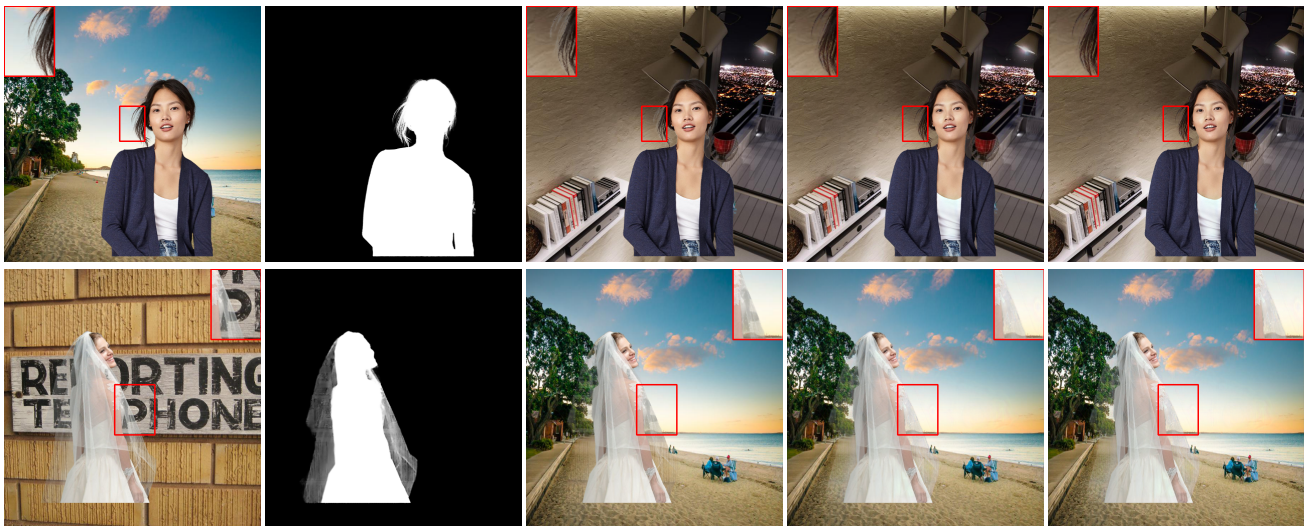

Figure 7: Improvement along blending boundaries. Compared with linear blending, our method obtains more seamlessly blending results, especially along boundaries and the lace regions.

Table 1: Quantitative evaluation with linear blending. We mainly compare the performance at the unknown regions of a fake trimap, which is obtained by erosion and dilation operations [10].

Methods	PSNR ↑	SSIM ↑
Linear blending	42.525	0.997
Ours	**43.798**	**0.998**

group-level consistency, for our network training and evaluation. We use 1356 groups for training, 100 group for validation, and 225 groups for testing. In each group, we extract the foreground from the the group's first half and take the second half as the target background. We additionally collect 1193 background photos from Internet and [15]. We split 900 backgrounds for training, 93 for validation, and 200 backgrounds for testing. To further demonstrate our performance on real-world photos, we also evaluate different approaches on the real-world harmonization benchmark, RealHM of [9]. For video harmonization evaluation, we additionally collect 12 portrait videos from the Internet and composite to our collected backgrounds.

Implementation details We train our seamless blending network and harmonization network separately. We apply random scaling and random rotation to the foregrounds and apply random shifts for compositing. We set the learning rate to 1e-4 and train the blending network for 80 epochs. We apply random flipping to train our harmonization network. We train our harmonization network with a learning rate 2e-4 for 114K steps. We resize the training pairs with resolution 512×512 for training both the blending network and harmonization network.

Table 2: Quantitative comparison with other methods and controlled experiments.

Methods	PSNR ↑	SSIM ↑
DoveNet-pre-trained [4]	25.09	0.94
RainNet-pre-trained [16]	24.71	0.93
SSH-pre-trained [9]	28.29	0.96
DoveNet [4]	31.09	0.95
RainNet [16]	**33.23**	**0.97**
Ours-norm-all	32.483	**0.97**
Ours-partial-background	32.65	**0.97**
Ours	**33.06**	**0.97**

5.2 Baselines and controlled experiments

We compare our method with following state-of-the-art harmonization models. We also design controlled experiments to verify the effectiveness of our proposed framework.

Linear blending with MODNet MODNet [10] is the state-of-the-art automatic human matting method. We use the pre-trained MODNet to extract alpha masks on our collected matting dataset. We then take the predicted mask, the source image, and the target background image as input, the composites with a perfect mask as ground-truth, to train our blending network. We utilize the ground truth alpha mask to identify the boundary region with erosion and dilation operations [10]. We mainly evaluate the effectiveness of our blending network at boundary regions with the linear blending method using the predicted mask.

DoveNet Cong *et al.* [4] publish a large-scale harmonization dataset through perturbing the foreground color of four public datasets that provide hard segmentation masks. They also propose

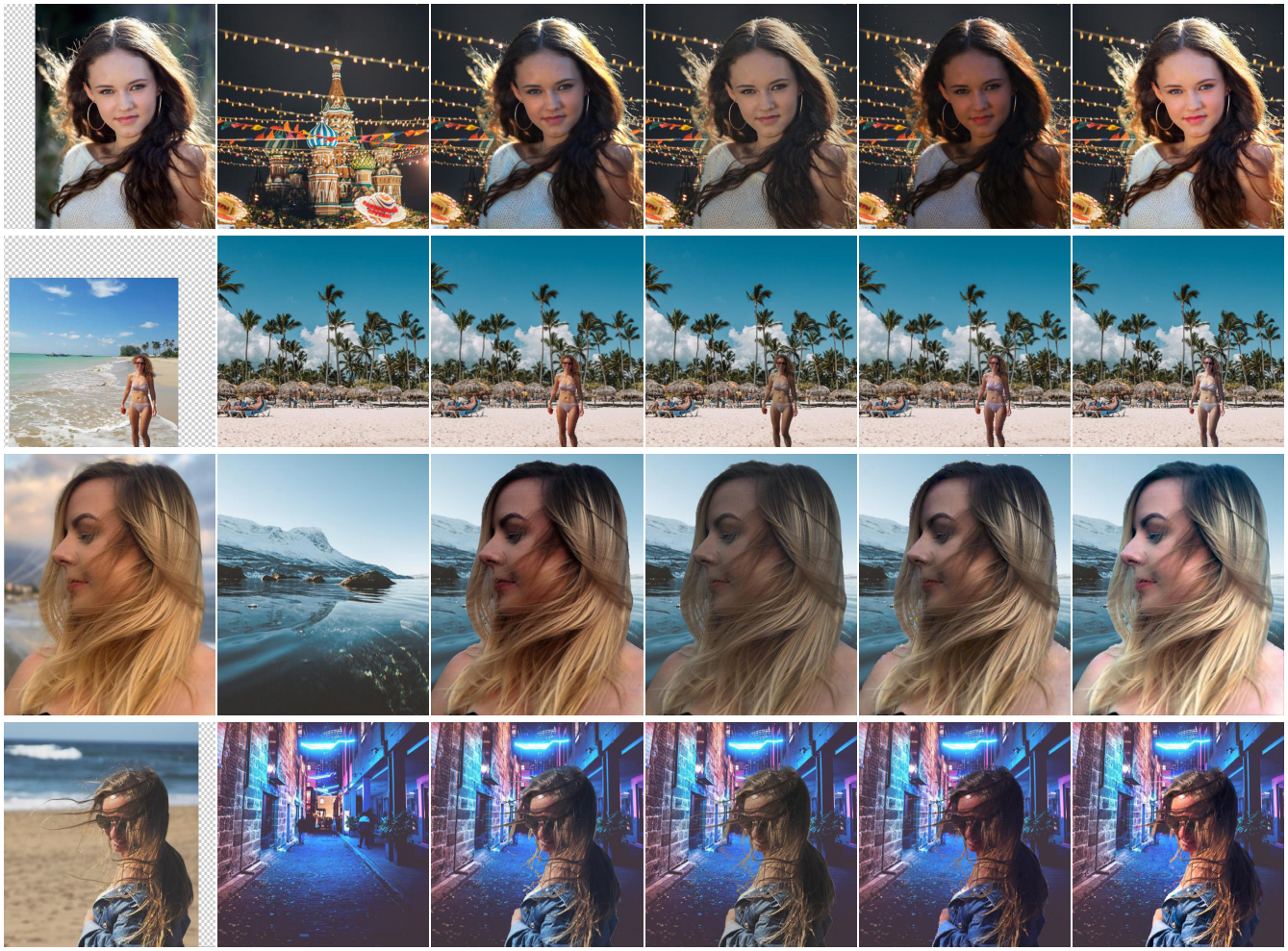

Figure 8: Comparisons between our framework and baselines. Our method can achieve more realistic harmonization results.

an image harmonization method with domain verification. We utilize their provided official pre-trained model for our evaluations.

RainNet Ling *et al.* [16] formulate the image harmonization method to a style transfer problem with their proposed region-aware AdaIN layers. We utilize their provided official pre-trained model for our evaluations.

SSH Different from previous image harmonization methods, Jiang *et al.* [9] propose a self-supervised framework to learn image harmonization. Different from ours, they still rely on linear blending to obtain the composites, which would produce noticeable artifacts along the boundaries.

Controlled experiments We conduct two ablation studies to our framework. First, we also utilize the backgrounds to normalize the foreground encoder features, denoted as **Ours-norm-all**. Second, we only extract the appearance feature from the partially occluded backgrounds, similarly as [16]. We denote this ablation as **Ours-partial-background**.

Table 3: User study results of preference rate of each method on the real-world benchmark RealHM [9].

Method	Copy&paste	DoveNet	SSH	RainNet	Ours
Ratio ↑	19%	16%	19%	21%	**25%**

5.3 Results

Quantitative results To quantitatively prove the effectiveness of our method, we use PSNR and SSIM to evaluate our blending and harmonization network. The comparisons with baselines are reported in Table 1 and Table 2. Compared with linear blending, our learnable blending network can produce much more accurate boundary blending results, which is proven by more than 1dB improvement in PSNR. On the other hand, our harmonization network can outperform SSH to a large extent in terms of PSNR and SSIM. We explain it as our network can see real-world photos

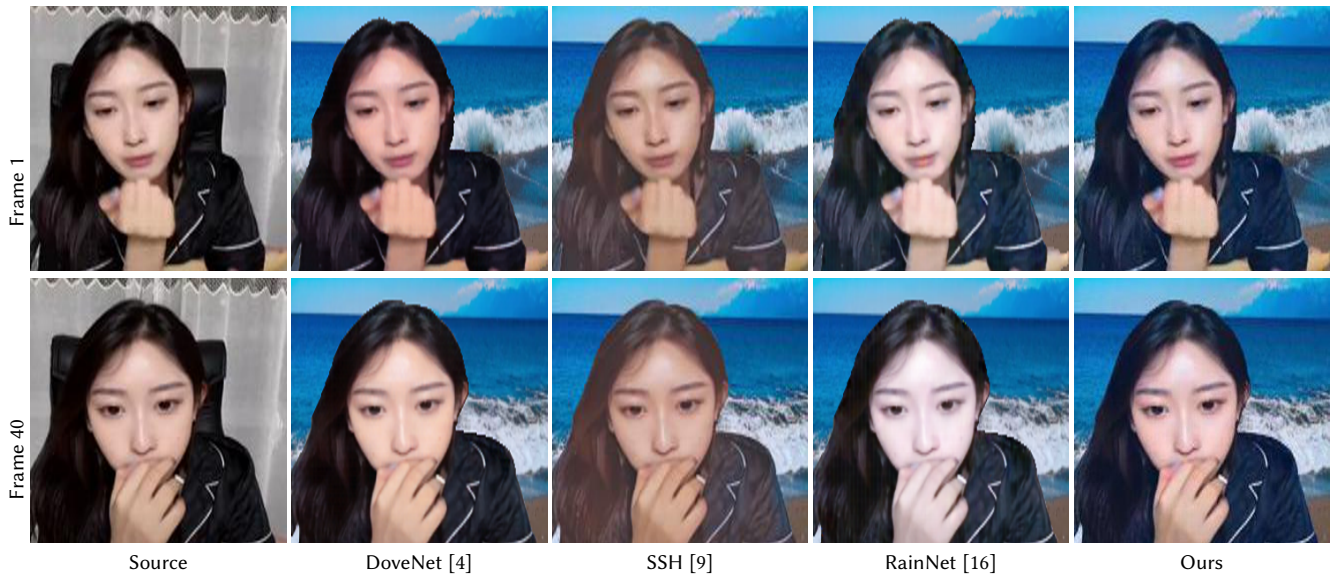

Figure 9: Extension to video harmonization. We show that our method can provide visually better and temporally more consistent video harmonization results than other methods.

during the training process and thus exceed simple color transfer regimes. Compared with DoveNet, our method can achieve good improvement in both PSNR and SSIM. This is because our network can utilize the entire background instead of the partial occludes ones, which limits the harmonization performance, especially when the foreground is huge. From the our ablation experiment **Ours-partial-background**, we also prove that complete background cue can help the image harmonization task. Compared with RainNet, we show our model can achieve comparable performance for harmonization. Moreover, prior harmonization networks ignore the blending process, which may produce unpleasant harmonization artifacts at boundary regions. Our method takes seamless blending as objective and produce more realistic composited photographs.

Qualitative results We present qualitative comparisons against baselines in Figure 8, Figure 7, and Figure 9. Note that we use the real-world harmonization benchmark RealHM [9] for comparing harmonization performance. In Figure. 7, we show that the linear blending will produce noticeable artifacts at the blending boundary. Our learnable blending, however, can seamlessly blend the foreground and background together, even at lace regions. In Figure 8, we show that SSH mainly produces color transfer from the target background to source images and overlooks the photorealism aspect of image harmonization, which results in unrealistic harmonization results compared with other methods. Compared with DoveNet and RainNet, our method can produce more harmonic results, which achieve a good balance between color saturation and brightness. In Figure 9, we show that our method can achieve more temporally consistent results on video harmonization. We explain it as our method can achieve a better content-style disentanglement than baselines through only inferring appearance from the complete backgrounds.

User study We conduct a user study to compare our results with baselines on a real-world benchmark provided by [9]. There are 198 portrait foreground/background pairs in the benchmark dataset. Each image is processed by five methods: copy-and-paste, DoveNet [4], RainNet [16], SSH [9], and ours. We split 198 images into 5 groups and assigned each group 5 persons to do the user study. We ask all the subjects to select the best results in each pair according to the factor: 1) the foreground should be harmonic with the background when considering color and illumination; 2) the composite images should be realistic without strange colors, especially at the skin regions. The results can be seen in Table 3. Interestingly, the copy-and-paste method has a similar preference rate with DoveNet and SSH. Our method has the highest preference rate among all the baselines.

6 CONCLUSION

We propose a unified pipeline to realistic image composition. Besides global color harmonization, we also regard seamless blending as another important objective for realistic photo composition, which is often overlooked by prior works. Our key insight is that complete target backgrounds play a vital role in both seamless blending and image harmonization. Instead of utilizing linear blending by prior works, we show that we can learn the blending process to remedy the boundary artifacts. Besides, we explicitly extract the appearance from the complete background to adjust the foreground features through AdaIN. We demonstrate that our proposed framework can produce both seamless blending and harmonic color adjustment for both images and videos. Our limitation mainly lies in the insufficiency of the existing matting datasets. If the large-scale matting dataset where real photos and alpha masks are provided, we can largely integrate our blending and harmonization network together to bring more compact models.

REFERENCES

[1] Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

[2] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. 2013. KNN matting. *PAMI* 35, 9 (2013), 2175–2188.

[3] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. 2016. Natural image matting using deep convolutional neural networks. Springer, 626–643.

[4] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. 2020. Dovenet: Deep image harmonization via domain verification. In *CVPR*.

[5] Xiaodong Cun and Chi-Man Pun. 2020. Improving the harmony of the composite image by spatial-separated attention module. *IEEE Transactions on Image Processing* 29 (2020), 4759–4771.

[6] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. 2021. Intrinsic Image Harmonization. In *CVPR*.

[7] Qiqi Hou and Feng Liu. 2019. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*.

[8] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*.

[9] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. 2021. SSH: A Self-Supervised Framework for Image Harmonization. In *ICCV*.

[10] Zhanghan Ke, Kaican Li, Yurou Zhou, Qiuhua Wu, Xiangyu Mao, Qiong Yan, and Rynson WH Lau. 2020. Is a Green Screen Really Necessary for Real-Time Portrait Matting? *arXiv preprint arXiv:2011.11961* (2020).

[11] Dingzeyu Li, Qifeng Chen, and Chi-Keung Tang. 2013. Motion-aware KNN Laplacian for video matting. 3599–3606.

[12] Jizhizi Li, Jing Zhang, and Dacheng Tao. 2021. Deep Automatic Natural Image Matting. *arXiv preprint arXiv:2107.07235* (2021).

[13] Yaoyi Li and Hongtao Lu. 2020. Natural image matting via guided contextual attention. In *AAAI*.

[14] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. 2021. PPR10K: A Large-Scale Portrait Photo Retouching Dataset with Human-Region Mask and Group-Level Consistency. In *CVPR*.

[15] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-time high-resolution background matting. In *CVPR*.

[16] Jun Ling, Han Xue, Li Song, Rong Xie, and Xiao Gu. 2021. Region-aware Adaptive Instance Normalization for Image Harmonization. In *CVPR*.

[17] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. 2020. Boosting semantic human matting with coarse annotations. In *CVPR*.

[18] Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson image editing. In *SIGGRAPH*. 313–318.

[19] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. 2020. Attention-guided hierarchical structure aggregation for image matting.

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

[21] Xuqian Ren, Yifan Li, and Chunlei Song. 2021. A Generative Adversarial Framework For Optimizing Image Matting And Harmonization Simultaneously. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1354–1358.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.

[23] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2020. Background matting: The world is your green screen.

[24] Ehsan Shahrian, Deepu Rajan, Brian Price, and Scott Cohen. 2013. Improving image matting using comprehensive sampling sets. In *CVPR*.

[25] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. 2016. Deep automatic portrait matting. In *ECCV*.

[26] Konstantin Sofiiuk, Polina Popenova, and Anton Konushin. 2021. Foreground-aware semantic representations for image harmonization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1620–1629.

[27] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. 2021. Semantic Image Matting.

[28] Kalyan Sunkavalli, Micah K Johnson, Wojciech Matusik, and Hanspeter Pfister. 2010. Multi-scale image harmonization. *ACM Transactions on Graphics* 29, 4 (2010), 1–10.

[29] Jingwei Tang, Yagiz Aksoy, Cengiz Oztireli, Markus Gross, and Tunc Ozan Aydin. 2019. Learning-based sampling for natural image matting. In *CVPR*.

[30] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep image harmonization. In *CVPR*.

[31] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. 2017. Deep image matting. In *CVPR*.

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

[33] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. 2021. Mask Guided Matting via Progressive Refinement Network. In *CVPR*.

[34] He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. 2021. Deep image compositing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 365–374.

[35] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. 2019. A late fusion cnn for digital matting. In *CVPR*.