Deep Image Compositing

He Zhang[*1], Jianming Zhang[*1], Federico Perazzi[*1], Zhe Lin[*1], Vishal M. Patel[*2]
1. Adobe Research 2. Johns Hopkins University

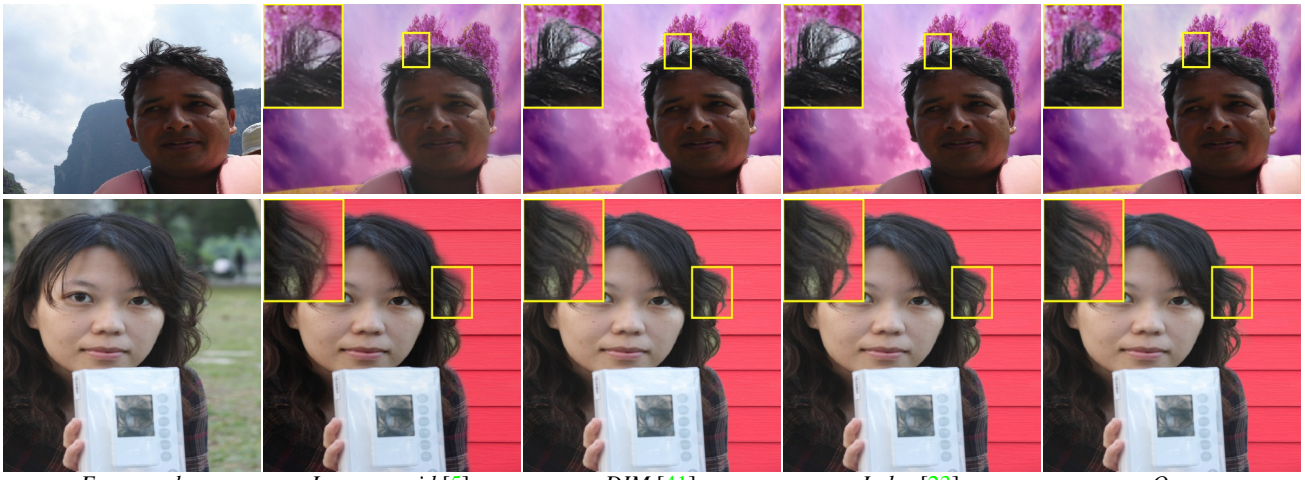


Figure 1: Portrait compositing results on the real-world images. All the methods are using the same estimated foreground mask. Previous methods suffer from problems such as halo artifacts and color contamination. Our method learns to generate better compositing results with less boundary artifacts and accurate foreground estimation. Best viewed in color.

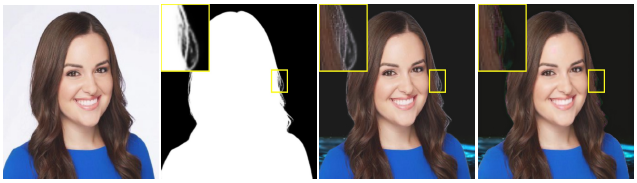
Abstract

Image compositing is a task of combining regions from different images to compose a new image. A common use case is background replacement of portrait images. To obtain high quality composites, professionals typically manually perform multiple editing steps such as segmentation, matting and foreground color decontamination, which is very time consuming even with sophisticated photo editing tools. In this paper, we propose a new method which can automatically generate high-quality image compositing without any user input. Our method can be trained end-to-end to optimize exploitation of contextual and color information of both foreground and background images, where the compositing quality is considered in the optimization. Specifically, inspired by Laplacian pyramid blending, a dense-connected multi-stream fusion network is proposed to effectively fuse the information from the foreground and background images at different scales. In addition, we introduce a self-taught strategy to progressively train from easy to complex cases to mitigate the lack of training data. Experiments show that the proposed method can automatically generate high-quality composites and outperforms existing methods both qualitatively and quantitatively.

1. Introduction

Image compositing is one of the most popular applications in image editing. A common scenario is to composite a portrait photo with a new background. Sample images for portrait compositing is shown in Fig. 1. To get high-quality composite images, professionals rely on image editing software to perform operations like segmentation, matting and foreground color decontamination. Although many parts of the workflow have been made relatively easier by software, it still requires a lot of expertise and manual efforts to create high-quality composited images. In this paper, we aim to fully automate the portrait image compositing process.

One straightforward solution is to use a salient object segmentation model [2, 46, 38, 9, 16, 25] to cut out the foreground region and then paste it on the target background image. However, such simple cut-and-paste approach with the segmentation mask usually results in undesirable artifacts along the object boundary. This is because pixels



Foreground *GT Mask* *Copy-Paste* *Our*

Figure 2: Leveraging high-quality masks for directly compositing (Copy-paste) may not result in high-quality compositing results. From the direct copy-paste results, it can be clearly observed that color-artifacts are along the boundary.

along the object boundary are usually linear combinations of both foreground and background. To address the boundary artifacts, previous approaches resort to low-level image blending methods such as Poisson blending [26], Laplacian blending [5], feathering, guided-filter [13], etc. However, these low-level blending methods often introduce other undesirable effects such as color distortion or non-smooth halo artifacts. Sample results are shown in Fig. 1.

A common solution to the boundary artifacts is to extract the object matte (*i.e.* alpha channel) from the foreground image using image matting methods [7, 41, 22, 29, 8, 10, 31, 1]. The ground truth matte controls the linear interpolation of foreground and background in the original input image. Hence, image mattes, if accurately predicted, are able to generate very convincing compositing results with natural blending along the boundary. However, the image matting problem is generally very challenging and it usually requires human input (eg trimap) to identify foreground, background and the uncertain regions to solve. In addition, mistakes in matting are not equally important to image compositing, and the matting methods cannot leverage that as they do not take the end compositing results into consideration. For example, as shown in Fig 2, though the high-quality ground truth mask (GT mask) is given to cut-out the foreground person and composite it onto a different background, yet the compositing result (Copy-paste result) contains obvious color artifacts along the boundary, which degrade the whole compositing quality.

In this work, we propose a deep learning based image compositing framework to directly generate a composited portrait image given a pair of foreground and background images. A foreground segmentation network together with a refinement network is employed to extract the portrait mask. Guided by the portrait mask, an end-to-end multi-stream Fusion (MLF) Network is proposed to merge information from both foreground and background images at different scales. The MLF network is inspired by the Laplacian Pyramid Blending method. It uses two encoders to extract different levels of feature maps for the foreground and background images respectively, and fuse them level-by-level through a decoder to reconstruct the final compositing result. To notice, the task of the harmonization [35, 34, 48] is different from ours. Their task is to harmonize the appearances (e.g. color) between the foreground and the background and they assume that an artifact-free mask is provided by the user. In contrast, our method is fully automatic and focuses on alleviating the boundary artifacts caused by imperfect foreground masking and color decontamination. Basically, our paper solves an orthogonal problem to color/appearance harmonization for image compositing. In addition, we propose an easy-to-hard self-taught based data augmentation scheme to generate high quality compositing data for training the MLF network. The basic idea is to use a MLF network, which is trained on simpler data, to composite more challenging training data for improving itself.

Experimental results evaluated on the synthetic images and real-world images demonstrate the effectiveness of the proposed method compared to previous methods. The superior perceptual quality of our method is validated though a user study. Sample results of our method can be found in Figs. 1 and 7 . To summarize, our contributions are

- an end-to-end deep learning based framework for fully automatic portrait image compositing,
- a novel multi-stream Fusion Image Compositing Network for fusing image features at different scales, and
- an easy-to-hard data-augmentation scheme for image compositing using self-taught.

2. Related Works

2.1. Image Compositing

Image compositing is a challenging problem in computer vision/ computer graphics, where salient objects from foreground image to be overlayed/composited onto given a background image. And the final goal of the image compositing is to generate realistic high-quality images. Many image editing applications fall into the category of the image compositing such as image harmonization [34, 35, 48, 32], image matting [1, 7, 8, 31, 10, 29, 41, 22, 37], image blending[5, 13, 26, 40].

The goal of classical image blending approaches is to guarantee that there is no apparent transition gap between source image and target image. Alpha blending [36] is the simplest and fastest method, but it blurs the fine details and may bring in halo-artifacts in the compositing images. To efficiently leverage the information from different scales, Burt and Adelson [5] proposed a multi-scale blending algorithm, named Laplacian pyramid blending. Similar idea has also been used for other low-level tasks such as image enhancement [20, 44] and Generative Adversarial Network (GAN) [11]. Alternatively, gradient-based approaches [26, 33, 18] also address this problem by adjusting the differences in color and illumination for the composited image globally.

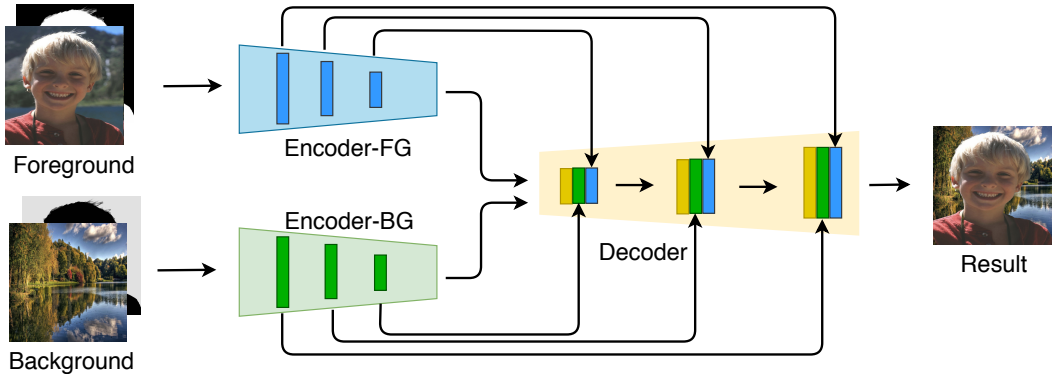


Figure 3: An overview of the proposed multi-stream Fusion Image Compositing Network, where features of different levels are extracted from foreground and background images separately, and then are fused together to generate high-quality compositing results. A pair of masks generated by segmentation models are used to guide the encoding process. See text for more details.

The most common compositing workflow is based on image matting. Matting refers to the process of extracting the alpha channel foreground object from an image. Traditional matting algorithms [1, 8, 41, 29, 23, 6] require a user defined trimap which limits their applications in automatic image process. Though recent works [31, 7] have leveraged the CNN models to automatically generate trimaps, they still regard trimap generation and alpha channel computation as two separated stages. In addition, the matting methods do not take the final compositing results into consideration. Instead, we propose an end-to-end image compositing framework that takes the final compositing performance as the optimization objective.

2.2. Data Augmentation

Data augmentation is a common technique to improve the training of deep neural networks. It helps to reduce over-fitting and improves the model generalization. Data-augmentation has been successfully applied to various computer vision applications both in low-level vision and high-level vision such as image enhancement [45, 43], image matting [41, 7], image harmonization [34], and object detection [12]. Most data augmentation methods are based on trivial transformations such as cropping, flipping, color shift, or adding noise to an image [12, 34, 49, 28]. In our problem, the training requires a triplet composed of a pair of foreground and background images and the target composited images, and traditional data augmentation methods cannot help to diversify the contents of the triplet. In this paper, we propose a self-taught method to automatically generate such triplet samples for our image compositing problem.

3. Deep Image Compositing

In this section, we present our Deep Image Compositing framework. Although we only implement it for portrait compositing in this paper, the formulation of the framework is general and we hope it can be useful in other image compositing applications, too.

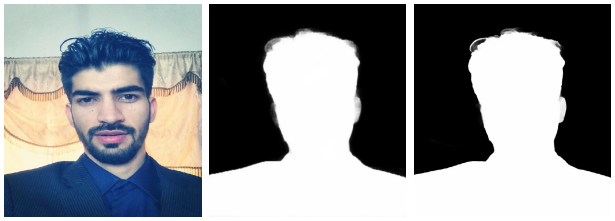
The proposed framework takes a pair of foreground and background images as input, and generates the composited image. It has three components: 1) a Foreground Segmentation Network, 2) a Mask Refinement Network and 3) a multi-stream Fusion Network. First, the segmentation network automatically extracts an object mask from the foreground image. Then, the mask refinement network takes the image and the mask as input to refine the mask boundary. After that, the refined mask, together with the foreground and background images, is passed to the multi-stream fusion network to generate the compositing results. We will describe these components as follows.

3.1. Multi-stream Fusion Network for Compositing

We present the multi-stream Fusion (MLF) Network first as it is independent of the other two components and can work with other segmentation and matting models, too. The goal of the MLF network is to provide natural blending between the foreground cutout and the background image, removing artifacts caused by color contamination, aliasing and the inaccuracy of the segmentation mask.

Our MLF network is inspired by the Laplacian pyramid method for image blending. The Laplacian blending [5] method computes image pyramid representations for both foreground and background images, and then blends different levels of details with varied softness along the mask boundary through the image pyramid representations. The final composited image can be reconstructed from the multi-stream fused image representations.

Similarly, the proposed MLF network consists of two separate encoders to extract the multi-scale features from foreground and background images separately. The input to both encoders is a concatenation of the image and a pre-computed soft mask. The mask for the foreground image is computed by our segmentation and refinement networks,



Foreground *Raw Mask* *Refined Mask*

Figure 4: Results on the segmentation mask estimation. It can be observed that the refined segmentation mask preserves better boundary details with more confidence.

and the mask for the background image is an inverted version of it. The foreground and background encoders then generate different levels of features, which correspond to the image pyramid representations in Laplacian blending.

At the end of the encoders, the highest-level of feature maps are concatenated and are passed to a decoder. At different decoding levels, feature maps get upsampled through deconvolution and are concatenated with the same level feature maps from the two encoders. At the end of the decoder, the composited image is reconstructed from the fused feature maps. This process is analogous to the fusion and reconstruction process of Laplacian pyramid blending. Dense-block [17] is leveraged as the basic building block for the discussed encoder and decoder architecture. Details of the proposed network is presented in the supplementary material. The overview of the MLF network is shown in Fig. 3.

As we can see, our MLF network can be regarded as an extension of the popular encoder-decoder network with short connections [27, 4]. Instead of a single-stream encoder in the previous models, our two-stream encoder pipeline encodes the foreground and background feature maps separately, which are fused later during the decoding process. We find that such two-stream design not only coincides with the Laplacian blending framework, but also provides better performance than the single-stream design in our experiment.

For training, the proposed method is optimized via both L1 loss and perceptual loss [19, 21] to encourage image-level perceptual realism [47] for the composited images. The loss function is defined as follows:

$$L_{\text{all}} = L_1 + \lambda_P L_P, \tag{1}$$

where L_1 denotes the L1 loss and L_P indicates the perceptual loss. Here, λ_P indicates the weights for the perceptual loss. The perceptual loss is evaluated on relu 1-1 and relu 2-1 layers of the pre-trained VGG [24] model.

3.2. Segmentation and Mask Refinement Networks

The foreground segmentation network can be implemented as a salient object segmentation model [2, 46, 38, 9, 16, 25] or specifically a portrait segmentation model [31, 7, 30], and we refer the readers to those related works for the details of model training and datasets. In our implementation, we use a salient object segmentation model [3] due to its speed and accuracy, but that our framework can work with any off-the-shelf salient object segmentation models.

However, the raw mask from the foreground segmentation model is often not very accurate at the object boundaries. The segmentation network also processes the image at low resolution, so the upsampled mask will further suffer from the up-sampling artifacts like jagged boundaries. Therefore, we propose a mask refinement network to refine the details along the object boundary and up-sample the mask with fewer artifacts.

The refinement network shares the same architecture as the segmentation network, except that input is a four-channel RGB-A image, where the fourth contains the raw segmentation mask. To make this mask refinement network focus on different levels of local details, we sample image patches of various sizes for the training. In training, a cropped version of the image and the pre-computed raw mask are passed to the refinement network to generate the local refined mask. The training uses the same data and and the same cross-entropy loss as used by the segmentation model. At testing, the refinement network takes the whole image and its mask as input.

The refinement network can be applied recursively at different scales. In our implementation, we first resize the image and the its raw mask to 320×320 and generate a refined mask at this resolution. Then we resize the image to 640×640 and upsample the refinemask to same size, and apply the refinement network again. We find this two-stage refinement scheme working very well in practice. A sample result of the refinement network is shown in Fig. 4. It can be observed that the refined mask preserves much better boundary details and reduce the fuzziness of the raw mask. This also makes the training of our fusion network easier.

4. Easy-to-Hard Data Augmentation

To train our multi-stream Fusion (MLF) network, each training sample is a triplet $[FG, BG, C]$, where FG is the foreground image, BG is the background image and C is the target composted image of FG and BG. As we want the MLF network to learn to produce a visually pleasing blending between FG and BG, the quality of the target image C is the key to our method. However, manually creating such high-quality compositing dataset requires expert-level human effort, which limits the scalability of the training data collection.

To address this issue and generate a relative large-scale image compositing dataset without much human annotation effort, we propose an easy-to-hard data-augmentation ap-

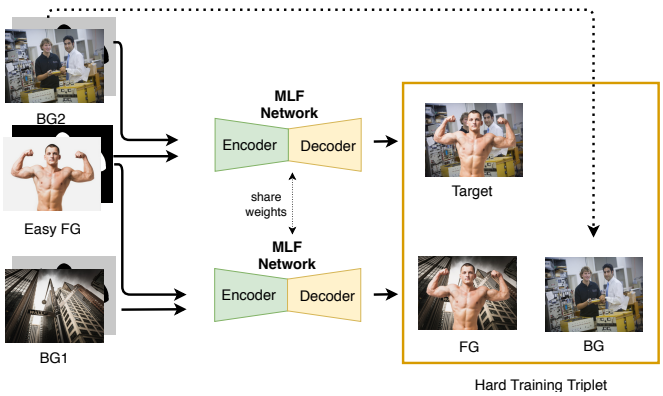


Figure 5: An overview of the proposed easy-to-hard data augmentation procedure.

Background *Foreground* *Target*

Figure 6: Samples of triplets generated by our self-taught data augmentation algorithm. It can be observed that the proposed data augmentation algorithm is able to generate high-quality with near-perfect target images.

proach using a self-taught scheme. The basic idea is to use the MLF network to generate more challenging data to improve itself. The MLF network is first trained on a few easy training triplets where the foreground images FG are all portrait images with simple color background. After that, we collect a lot of such simple portrait images and use the MLF network to generate more challenging training triplets for the next training stage. The overview of this data augmentation scheme is shown in Fig. 5 and we describe more details below.

We first use a small matting dataset [41] to create a simple compositing training set. Images in the matting dataset have alpha channel and were processed by color decontamination. Thus, they can be composited to any background images using the alpha channel. To generate a easy training triplet, the foreground images FG is generated by compositing the matting image with pure color background; the background image BG can be a random web image; and the target image C is created by using the alpha channel of the matting image as well. By firstly training our MLF network on these triplets, the network learns to blend an easy foregrond image onto a random background image.

We then use our specifically trained MLF network to generate harder training triplets. We collect a lot of web portrait images with simple background, with which we generate composited images with random background images using the MLF network (see Fig. 5). Given a simple portrait image, which is denoted as Easy FG in Fig. 5, we sample two random background images $BG1$ and $BG2$ to generate two composited images using the MLF network. Without loss of generality, the composited image of $BG2$ and Easy FG is used as a new target image C'; the other composited image then becomes a new foreground image FG', and $BG2$ will be used as the new background image BG'. As we can see from Fig. 5, the new triplet $[FG', BG', C']$ follows the compositing relationship

$$C' = \text{Easy } FG \oplus BG2 = FG' \oplus BG', \qquad (2)$$

whtere Easy FG and FG' share the same foreground content, and $BG' = BG2$. The compositing operation is denoted as $\oplus$.

In this way, we generated high-quality hard triplets to augment the original matting training set. Sample triplets are shown in Fig. 6. It can be observed that the proposed data augmentation algorithm is able to generate high-quality compositing targets. Our results in the next section show that these self-generated training sample is essential for the good performance of our method.

5. Experiments

We evaluate our deep image compositing method through quantitative and qualitative evaluations. A user study is also performed to measure the users' preference regarding the perceptual quality of the compositing results. Finally, we perform some ablation studies to validate the main components of our method.

Datasets: The segmentation and refinement network is trained on the DUTS [39, 42], MSRA-10K [15, 14] and Portrait segmentation [31, 30] datasets . The multi-stream fusion compositing network is trained using the synthesized dataset via the proposed self-taught data augmentation method together with a matting-based compositing dataset. Similar to [41], the matting-based compositing dataset is composed of 30000 training images generated by compositing foreground images to random background images using the ground truth mattes. In addition, we also synthesize a testing dataset using the proposed self-taught data-augmentation method, denoted as **SynTest**. SynTest is used for quantitative evaluations. We leverage the PSNR as the measurements to measure the final compositing quality.

Implementation Details: The segmentation and refinement module is optimized via ADAM algorithm with a learning rate of 2×10^{-3} and a batch size of 8. All the training samples for the segmentation and refinement modules are resized to 256×256.

Similarly, the multi-stream fusion compositing network is trained using ADAM with a learning rate of 2×10^{-3} and batch size of 1. All the training samples are cropped and

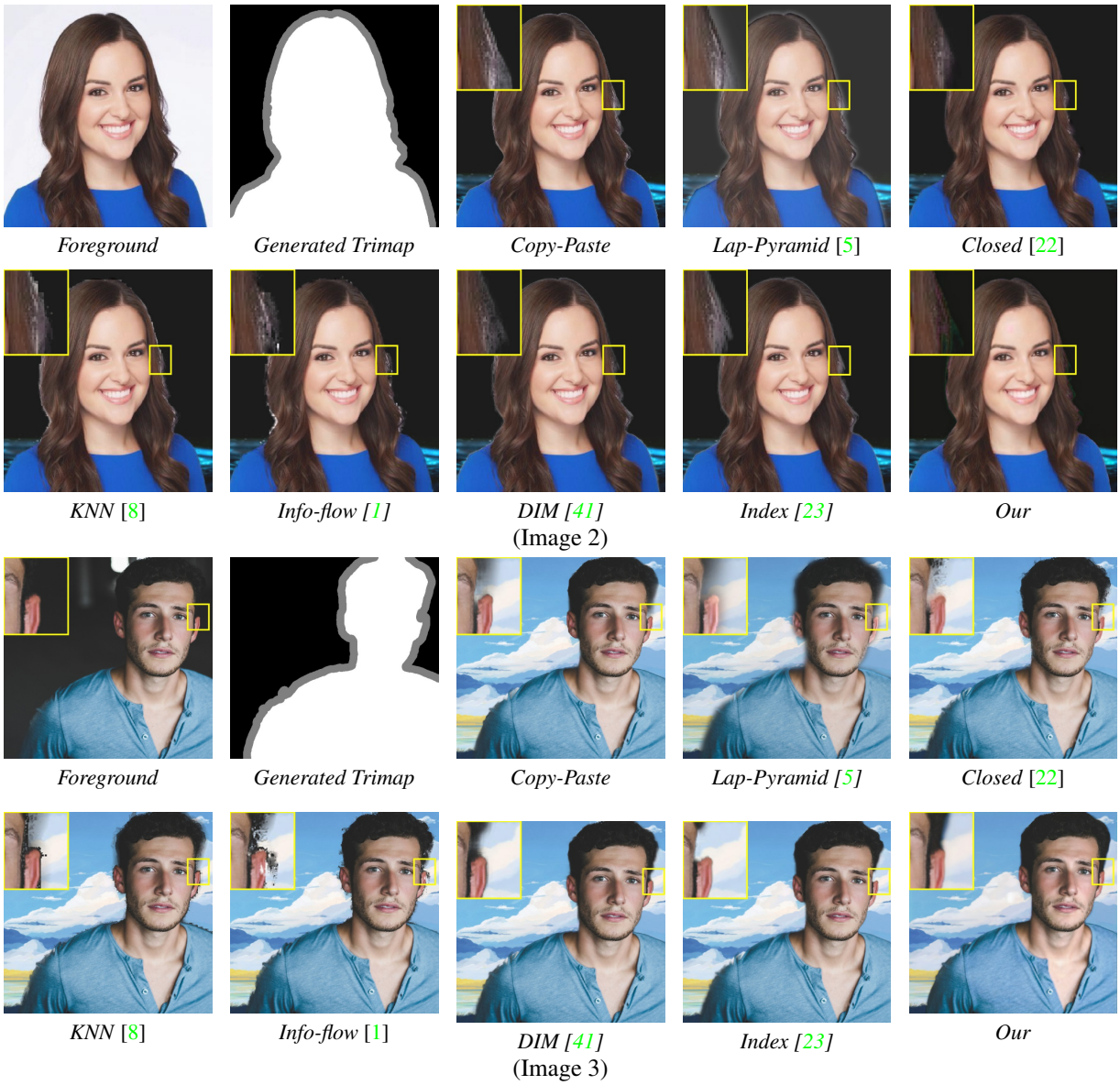


Figure 7: Results compared with other methods.

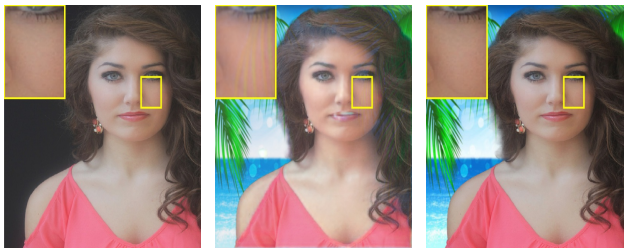
Table 1: Quantitative results evaluated on **SynTest** compared with the other methods. Evaluation on unknown regions.

	Copy-paste	Lap-Pyramid [5]	Closed [22]	KNN [8]	Info-flow [1]	DIM [41]	Index [23]	Our
PSNR (dB)	17.88	17.85	18.23	18.21	17.88	18.40	19.01	**19.34**

resized to 384 × 384 and all the testing samples are resized to 768 × 768. We trained the network for 200000 iterations and choose $\lambda_P = 0.8$ for the loss in Eqn. 1. Details of the proposed multi-stream fusion network is discussed in the supplementary material.

Compared Methods: In this paper, the proposed method is compared with traditional blending-based compositing methods such as Laplacian pyramid blending [5]. We also evaluate the matting-based compositing approach using state-of-the-art matting methods such as Closed-Form (Closed) [22], KNN [8], Information-Flow (Info-flow) [1], Deep Image Matting (DIM) [41] and Index-net [23]. In addition, we also compare one baseline method called copy-paste. For copy-paste, the refined segmentation mask estimated from the refinement segmentation module is used as the soft alpha matte for the compositing.

For fair comparison, all the compared methods use the same refined mask as our method. For the feathering method, we apply Gaussian blur with $\sigma = 2$ to soften the mask. For the Laplacian pyramid blending [5], we use the OpenCV implementation. As matting-based methods require trimaps, we binarize the refined mask and then gener-



Foreground *w/o-DataAug* *Our*

Figure 8: Results on the Ablation 1 on a real-world image. Without data augmentation, the baseline MLF network often makes the foreground region mixed with the background colors, leading to color shift artifact.

ate a pseudo-trimap by labeling a narrow boundary band of width 16 as unknown. Sample trimaps are shown in Fig. 7. To notice, an automatic color-decontamination algorithm [22] is used for the matting-based compositing methods to enhance their compositing quality.

5.1. Results

As our task is focusing more on perceptual quality, our evaluation is composed of subject judgment and quantitative evaluation. The quantitative evaluation uses two standard metrics PSNR to demonstrate the compositing quality, and it serves as a verification process. We do not deliberately pursue high scores of these metrics.

Some visual comparisons are shown in Fig. 7[1]. It can be seen that the Feathering and Laplacian pyramid methods are able to smooth the hard mask and improve the visual quality along the boundary. However, these two blending methods bring in halo artifacts and the object boundaries tend to be over-smoothed in the composited images. In contrast, the matting-based methods with color decontamination are able to generate compositing results with fine details along the object boundary, but sometimes the boundary artifacts becomes more obvious when the matting is not successful on challenging scenarios, as shown in the third image in Fig. 7. In addition, the imperfect trimap generated using the refined segmentation mask may also introduce more difficulty for the matting-based methods, and we find some of matting methods are sensitive to the choice of trimaps. Overall, our method is much more robust to the mistakes in the segmentation mask, and is able to generate higher quality compositing results in most of the cases. It can preserve natural details along the hair and object boundaries, and is also able to complete the missing part that is not completely captured by the input segmentation mask. Quantitative results evaluated on the synthetic data also demonstrate the effectiveness of the proposed methods, as shown in Table 1. The quantitative results in Table 1 are computed on the unknown regions only. In this setting, our method outperforms Deep Image Matting (DIM) by nearly 1 dB in PSNR and state-of-the-art Index-net [23] by 0.3 dB. Quantitative results evaluated on the whole image region is shown in supplementary material. Still, the numbers can only partially convey the performance gain of our method.

User Study: To further evaluate the perceptual quality of our results, we perform a user study. In this user study, we compare our results with the Laplacian Pyramid Blending method [5], and matting based methods using Closed-form [22], Information-flow [1] DIM [41] and Index [23]. We also include a Single-Stream network (Single-Enc) baseline to verify the architecture design of the MLF network (See the Ablation 2 in Sec 5.2 for more details). The Copy-paste baseline with refined soft mask is also included to demonstrate the advantage of our deep image composting framework.

This study involves 44 participants including Photoshop experts. During this study, each participant was shown 14 image sets, each consisting of the foreground images and compositing results of all compared methods. All testing images and compositing results are included in the supplementary materials. In each image set, participants were asked to pick and rank the favorite 3 results. The compostsing results in each set is randomly shuffled to avoid selection bias. We report the average ranking (lower the better) in Table 2. Compositing results that are not selected are assigned a rank score of 8 for more penalty.

Our method achieves the best ranking score. Among the 14 test samples, our method ranks the first on 9 images. The runner-up, which is Index-net [23], ranks the first on 4 images. And DIM [41] is the third place, which ranks the first on 1 image. Other matting-based methods are generally ranked lower than the our baselines. One major reason is that they often produce color artifacts along the object boundary, especially on challenging images where color contrast between object and background is not strong enough, or there are strong textures on the background near the object boundary. Moreover, some matting methods are sensitive to the trimaps and their performance may degrade significantly when the trimap is not accurate enough. In such cases, the users even prefer the smoother Lap-Pyramid results over the sharper matting-based results. These findings suggest the necessity of an end-to-end formulation for the image compositing problem.

5.2. Ablation Study

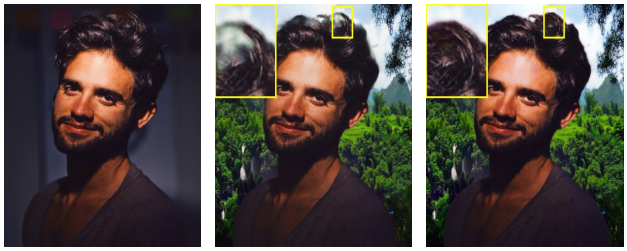
We conduct three ablation studies to demonstrate the effectiveness of the proposed method. The quantitative evaluation is performed on the synthesized **SynTest** datasets.

Ablation 1: Effectiveness of Data Augmentation. We evaluate the effectiveness of our self-teach data augmentation method. We train a baseline MLF network only on the matting-based compositing data (see Sec. 5) without using

[1]Due to the page limit, feathering results are put into the supplementary material.

Table 2: User-study results compared with the other methods. (Lower is the better.)

	Lap-Pyramid [5]	Closed [22]	Info-flow [1]	DIM [41]	Index [23]	Copy-paste	Single-Enc	Our
Average rank	5.02	4.81	4.50	3.17	3.69	4.22	4.15	**3.05**



Foreground *Single-Enc* *Our*

Figure 9: Results of the Ablation 2 on a real-world image. The baseline using a single-stream encoder tends to have issues in preserving the foreground regions.

Table 3: Quantitative results on **SynTest** of *Ablation 1-3*, where w/o-DataAug denotes the network trained without our data augmention, Single-Enc denotes a network with a single-stream encoder and w/o-RefNet means the baseline without the segmentation refinement network. Evaluation on unknown regions only.

	w/o-DataAug	Single-Enc	w/o-RefNet	Our
PSNR (dB)	18.00	19.05	18.22	**19.34**

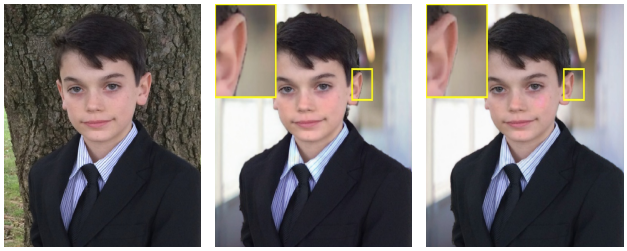
the self-taught based data augmentation (denoted as w/so-DataAug).

By visually checking testing samples, we find that the baseline MLF network is much less robust without the self-taught data augmentation. In many cases, the foreground tend be transparent, leading to color shift on the foreground. A sample result is shown in Fig. 8. Quantitative results in Table 3 also verify this observation. It shows that large training data is essential for training a robust MLF network and our self-taught data augmentation can effectively mitigate the lack of training data.

Ablation 2: Effectiveness of the Two-stream Encoder. Next, we demonstrate the effectiveness of our two-stream encoder design. We train a baseline network with a single-stream encoder-decoder structure (denoted as Single-Enc), where the foreground and background image, together with the refined mask, are concatenated as the input to the network. The backbone model of this baseline is the same as our full model. We do make sure the parameters for both single and two-stream have approximately the same number of parameters. We increase the number of channels for the encoder in the single-stream network

Similar to the Ablation 1, we evaluate this baseline on both the synthetic datasets and real-world images. A visual comparison is shown in Fig. 9. It can be observed from the zoom-in region that the single-stream network is less successful in preserving the foreground regions and causes more artifacts along the boundary. Quantitative results in Table 3 is also consistent this observation.

Ablation 3: Effectiveness of the Mask Refinement Network. We further evaluate the benefit of introducing the mask refinement network. For this baseline (denoted as w/o-RefNet), we remove the refinement network and directly use the raw segmentation mask for testing. To make it fair, we also re-train the MLF network using the raw mask.



Foreground *w/o-RefNet* *Our*

Figure 10: Results of the Ablation 3 on a real-world image. It can be observed from the zoom-in region that the refinement networks enables cleaner boundaries in the composited image.

Sample visual result is shown in Fig. 10. In general, the baseline is able to generate high-quality compositing results, but the object boundaries often contain contaminated colors that belong to the original background of the portrait photo. Quantitative results in Table 3 echos this observation.

Failure Case: The proposed MLF compoting network is robust to small errors in the segmentation mask, but it still relies on the general quality of the masks, as indicated in Ablation 3. Most of our failure cases are caused by failures of the segmentation network. Sample failure cases are shown in the supplementary material.

6. Conclusion

In this paper, we propose an end-to-end image compositing framework, where a saliency segmentation model with a refinement module is embedded into the network. To efficiently leverage features of both foreground and background from different scales, a multi-stream fusion network is proposed to generate the final compositing results. Furthermore, a self-taught data-augmentation algorithm is leveraged to augment the current compositing datasets. Experiments evaluated on both synthetic images and real-world images demonstrate the effectiveness of the proposed method compared to other methods. The user-study also show that the proposed method is able to generate better compositing results with good perceptual quality.

References

[1] Y. Aksoy, T. Ozan Aydin, and M. Pollefeys. Designing effective inter-pixel information flow for natural image matting.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 29–37, 2017.

[2] M. Amirul Islam, M. Kalash, and N. D. Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7142–7150, 2018.

[3] Anonymous. Deep j-net: A neural network architecture for accurate salient image segmentation.

[4] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[5] P. J. Burt and E. H. Adelson. A multiresolution spline with application to image mosaics. *ACM transactions on Graphics*, 2(4):217–236, 1983.

[6] S. Cai, X. Zhang, H. Fan, H. Huang, J. Liu, J. Liu, J. Liu, J. Wang, and J. Sun. Disentangled image matting, 2019.

[7] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai. Semantic human matting. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 618–626. ACM, 2018.

[8] Q. Chen, D. Li, and C.-K. Tang. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

[9] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[10] D. Cho, Y.-W. Tai, and I. Kweon. Natural image matting using deep convolutional neural networks. In *European Conference on Computer Vision*, pages 626–643. Springer, 2016.

[11] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In *29th Annual Conference on Neural Information Processing Systems, NIPS 2015*, pages 1486–1494. Neural information processing systems foundation, 2015.

[12] D. Dwibedi, I. Misra, and M. Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1301–1310, 2017.

[13] K. He, J. Sun, and X. Tang. Guided image filtering. In *European conference on computer vision*, pages 1–14. Springer, 2010.

[14] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In *IEEE CVPR*, pages 3203–3212, 2017.

[15] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 41(4):815–828, 2019.

[16] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

[17] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten. Densely connected convolutional networks. *arXiv preprint arXiv:1608.06993*, 2016.

[18] J. Jia, J. Sun, C.-K. Tang, and H.-Y. Shum. Drag-and-drop pasting. *ACM Transactions on Graphics (TOG)*, 25(3):631–637, 2006.

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[20] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 624–632, 2017.

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[22] A. Levin, D. Lischinski, and Y. Weiss. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2008.

[23] H. Lu, Y. Dai, C. Shen, and S. Xu. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3266–3275, 2019.

[24] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[25] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, pages 733–740. IEEE, 2012.

[26] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003.

[27] O. Ronneberger, P. Fischer, and T. 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.

[28] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[29] E. Shahrian, D. Rajan, B. Price, and S. Cohen. Improving image matting using comprehensive sampling sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 636–643, 2013.

[30] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs. Automatic portrait segmentation for image stylization. In *Computer Graphics Forum*, volume 35, pages 93–102. Wiley Online Library, 2016.

[31] X. Shen, X. Tao, H. Gao, C. Zhou, and J. Jia. Deep automatic portrait matting. In *European Conference on Computer Vision*, pages 92–107. Springer, 2016.

[32] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister. Multi-scale image harmonization. In *ACM Transactions on Graphics (TOG)*, volume 29, page 125. ACM, 2010.

[33] R. Szeliski, M. Uyttendaele, and D. Steedly. Fast poisson blending using multi-splines. In *2011 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2011.

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

[35] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, and M.-H. Yang. Sky is not the limit: semantic-aware sky replacement.

[36] M. Uyttendaele, A. Eden, and R. Skeliski. Eliminating ghosting and exposure artifacts in image mosaics. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001.

[37] J. Wang, M. F. Cohen, et al. Image and video matting: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(2):97–175, 2008.

[38] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *European conference on computer vision*, pages 825–841. Springer, 2016.

[39] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *European conference on computer vision*, pages 825–841. Springer, 2016.

[40] H. Wu, S. Zheng, J. Zhang, and K. Huang. Gp-gan: Towards realistic high-resolution image blending. *arXiv preprint arXiv:1703.07195*, 2017.

[41] N. Xu, B. Price, S. Cohen, and T. Huang. Deep image matting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2017.

[42] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[43] R. Yasarla and V. M. Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8405–8414, 2019.

[44] H. Zhang and V. M. Patel. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3194–3203, 2018.

[45] H. Zhang and V. M. Patel. Density-aware single image deraining using a multi-stream dense network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 695–704, 2018.

[46] J. Zhang and S. Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013.

[47] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. 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.

[48] J.-Y. Zhu, P. Krahenbuhl, E. Shechtman, and A. 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.

[49] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. D. Newsam, A. Tao, and B. Catanzaro. Improving semantic segmentation via video propagation and label relaxation. *CoRR*, abs/1812.01593, 2018.