Deep Image Harmonization

Yi-Hsuan Tsai[1] Xiaohui Shen[2] Zhe Lin[2] Kalyan Sunkavalli[2] Xin Lu[2] Ming-Hsuan Yang[1]
[1]University of California, Merced [2]Adobe Research
[1]{ytsai2, mhyang}@ucmerced.edu [2]{xshen, zlin, sunkaval, xinl}@adobe.com

Abstract

Compositing is one of the most common operations in photo editing. To generate realistic composites, the appearances of foreground and background need to be adjusted to make them compatible. Previous approaches to harmonize composites have focused on learning statistical relationships between hand-crafted appearance features of the foreground and background, which is unreliable especially when the contents in the two layers are vastly different. In this work, we propose an end-to-end deep convolutional neural network for image harmonization, which can capture both the context and semantic information of the composite images during harmonization. We also introduce an efficient way to collect large-scale and high-quality training data that can facilitate the training process. Experiments on the synthesized dataset and real composite images show that the proposed network outperforms previous state-of- the-art methods.

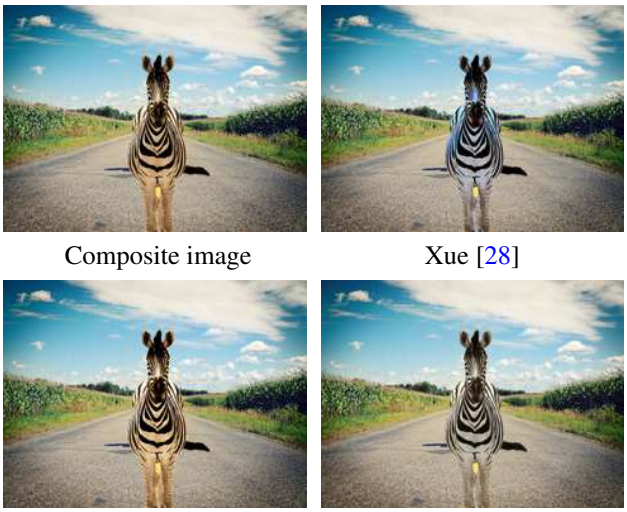


Figure 1. Our method can adjust the appearances of the composite foreground to make it compatible with the background region. Given a composite image, we show the harmonized images generated by [28], [32] and our deep harmonization network.

1. Introduction

Compositing is one of the most common operations in image editing. To generate a composite image, a foreground region in one image is extracted and combined with the background of another image. However, the appearances of the extracted foreground region may not be consistent with the new background, making the composite image unrealistic. Therefore, it is essential to adjust the appearances of the foreground region to make it compatible with the new background (Figure 1). Previous techniques improve the realism of composite images by transferring statistics of hand-crafted features, including color [13, 28] and texture [25], between the foreground and background regions. However, these techniques do not take the contents of the composite images into account, leading to unreliable results when appearances of the foreground and background regions are vastly different.

In this work, we propose a learning-based method by training an end-to-end deep convolutional neural network (CNN) for image harmonization, which can capture both the context and semantic information of the composite images during harmonization. Given a composite image and a foreground mask as the input, our model directly outputs a harmonized image, where the contents are the same as the input but with adjusted appearances on the foreground region. Context information has been utilized in several image editing tasks, such as image enhancement [6, 29], image editing [27] and image inpainting [20]. For image harmonization, it is critical to understand what it looks like in the surrounding background region near the foreground region. Hence foreground appearances can be adjusted accordingly to generate a realistic composite image. Toward this end, we train a deep CNN model that consists of an encoder to capture the context of the input image and a decoder to reconstruct the harmonized image using the learned representations from the encoder.

In addition, semantic information is of great importance

to improve image harmonization. For instance, if we know the foreground region to be harmonized is a sky, it is natural to adjust the appearance and color to be blended with the surrounding contents, instead of making the sky green or yellow. However, the above-mentioned encoder-decoder does not explicitly model semantic information without the supervision of high-level semantic labels. Hence, we incorporate another decoder to provide scene parsing of the input image, while sharing the same encoder for learning feature representations. A joint training scheme is adopted to propagate the semantic information to the harmonization decoder. With such semantic guidance, the harmonization process not only captures the image context but also understands semantic cues to better adjust the foreground region.

Training an end-to-end deep CNN requires a large-scale training set including various and high-quality samples. However, unlike other image editing tasks such as image colorization [30] and inpainting [20] where unlimited amount of training data can be easily generated, it is relatively difficult to collect a large-scale training set for image harmonization, as generating composite images and ground truth harmonized output requires professional editing skills and a considerable amount of time. To solve this problem, we develop a training data generation method that can synthesize large-scale and high-quality training pairs, which facilitates the learning process.

To evaluate the proposed algorithm, we conduct extensive experiments on synthesized and real composite images. We first quantitatively compare our method with different settings to other existing approaches for image harmonization on our synthesized dataset, where the ground truth images are provided. We then perform a user study on real composite images and show that our model trained on the synthesized dataset performs favorably in real cases.

The contributions of this work are as follows. First, to the best of our knowledge, this is the first attempt to have an end-to-end learning approach for image harmonization. Second, we demonstrate that our joint CNN model can effectively capture context and semantic information, and can be efficiently trained for both the harmonization and scene parsing tasks. Third, an efficient method to collect large-scale and high-quality training images is developed to facilitate the learning process for image harmonization.

2. Related Work

Our goal is to harmonize a composite image by adjusting its foreground appearances while keeping the same background region. In this section, we discuss existing methods closely related to this setting. In addition, the proposed method adopts a learning-based framework and a joint training scheme. Hence recent image editing methods within this scope are also discussed.

Image Harmonization. Generating realistic composite images requires a good match for both the appearances and contents between foreground and background regions. Existing methods use color and tone matching techniques to ensure consistent appearances, such as transferring global statistics [24, 23], applying gradient domain methods [21, 26], matching multi-scale statistics [25] or utilizing semantic information [27]. While these methods directly match appearances to generate realistic composite images, realism of the image is not considered. Lalonde and Efros [13] predict the realism of photos by learning color statistics from natural images and use these statistics to adjust foreground appearances to improve the chromatic compatibility. On the other hand, a data-driven method [10] is developed to improve the realism of computer-generated images by retrieving a set of real images with similar global layouts for transferring appearances.

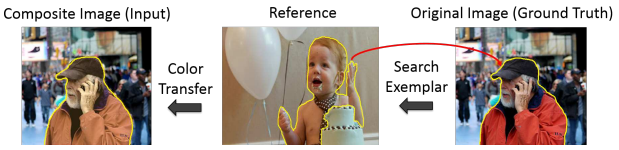
In addition, realism of the image has been studied and used to improve the harmonization results. Xue et al. [28] perform human subject experiments to identify most significant statistical measures that determine the realism of composite images and adjust foreground appearances accordingly. Recently, Zhu et al. [32] learn a CNN model to predict the realism of a composite image and incorporate the realism score into a color optimization function for appearance adjustment on the foreground region. Different from the above-mentioned methods, our end-to-end CNN model directly learn from pairs of a composite image as the input and a ground truth image, which ensures the realism of the output results.

Learning-based Image Editing. Recently, neural network based methods for image editing tasks such as image colorization [7, 14, 30], inpainting [20] and filtering [18], have drawn much attention due to their efficiency and impressive results. Similar to autoencoders [1], these methods adopt an unsupervised learning scheme that learns feature representations of the input image, where raw data is used for supervision. Although our method shares the similar concept, to the best of our knowledge it is the first end-to-end trainable CNN architecture designed for image harmonization.

However, these image editing pipelines may suffer from missing semantic information in the finer level during reconstruction, and such semantics are important cues for understanding image contents. Unlike previous methods that do not explicitly use semantics, we incorporate an additional model to predict pixel-wise scene parsing results and then propagate this information to the harmonization model, where the entire framework is still end-to-end trainable.

3. Deep Image Harmonization

In this section, we describe the details of our proposed end-to-end CNN model for image harmonization. Given



(a) Miscrosoft COCO & Flickr

(b) MIT-Adobe FiveK

Figure 2. Data acquisition methods. We illustrate the approaches for collecting training pairs for the datasets (a) Miscrosoft COCO and Flickr via color transfer, and (b) MIT-Adobe FiveK with different styles.

a composite image and a foreground mask as the input, our model outputs a harmonized image by adjusting foreground appearances while retaining the background region. Furthermore, we design a joint training process with scene parsing to understand image semantics and thus improve harmonization results. Figure 3 shows an overview of the proposed CNN architecture. Before describing this network, we first introduce a data collection method that allows us to obtain large-scale and high-quality training pairs.

3.1. Data Acquisition

Data acquisition is an essential step to successfully train a CNN. As described above, an image pair containing the composite and harmonized images is required as the input and ground truth for the network. Unlike other unsupervised learning tasks such as [30, 20] that can easily obtain training pairs, image harmonization task requires expertise to generate a high-quality harmonized image from a composite image, which is not feasible to collect large-scale training data.

To address this issue, we start from a real image which we treat as the output ground truth of our network. We then select a region (e.g., an object or a scene) and edit its appearances to generate an edited image which we use as the input composite image to the network. The overall process is described in Figure 2. This data acquisition method ensures that the ground truth images are always realistic so that the goal of the proposed CNN is to directly reconstruct a realistic output from a composite image. In the following, we introduce the details of how we generate our synthesized dataset.

Images with Segmentation Masks. We first use the Microsoft COCO dataset [17], where the object segmentation masks are provided for each image. To generate synthesized composite images, we randomly select an object and edit its appearances via a color transfer method. In order to ensure that the edited images are neither arbitrary nor unrealistic in color and tone, we construct the color transfer functions by searching for proper reference objects.

Table 1. Number of training and test images on three synthesized datasets.

	MSCOCO	MIT-Adobe	Flickr
Training set	51187	4086	4720
Test set	3842	68	96

Specifically, given a target image and its corresponding object mask, we search a reference image which contains the object with the same semantics. We then transfer the appearance from the reference object to the target object. As such, we ensure that the edited object still looks plausible but does not match the background context. For color transfer, we compute statistics of the luminance and color temperature, and use the histogram matching method [16].

To generate a larger variety of transferred results, we apply different transfer parameters for both the luminance and color temperature on one image, so that our learned network can adapt to different scenarios in real cases. In addition, we apply an aesthetics prediction model [11] to filter out low-quality images. An example of generated synthesized input and output pairs are shown in Figure 2(a).

Images with Different Styles. Although the Microsoft COCO dataset provides us with rich object categories, it is still limited to certain objects. To cover more object categories, we augment it with the MIT-Adobe FiveK dataset [3]. In this dataset, each original image has another 5 different styles that are re-touched by professional photographers using Adobe Lightroom, resulting in 6 editions of the same image. To edit the original image, we begin with one randomly selected style and manually segment a region. We then crop this segmented region and overlay on the image with another style to generate the synthesized composite image. An example set is presented in Figure 2(b).

Flickr Images with Diversity. Since images in the MIT-Adobe FiveK and Microsoft COCO datasets only contain certain scenes and styles, we collect a dataset from Flickr with larger diversity such as images containing different scenes or stylized images. To generate input and ground truth pairs, we apply the same color transfer technique described for the Microsoft COCO dataset. However, since there is no semantic information provided in this dataset to search proper reference objects for transfer, we use a pre-trained scene parsing model [31] to predict semantic pixel-wise labels. We then compute a spatial-pyramid label histogram [15] of the target image and retrieve reference images from the ADE20K dataset [31] with similar histograms computed from the ground truth annotations.

Next, we manually segment a region (e.g., an object or

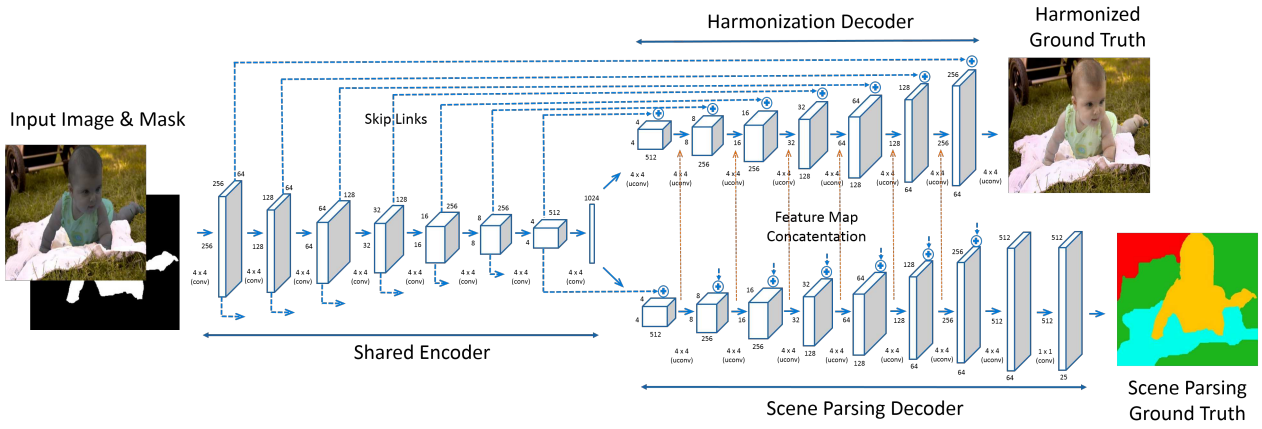


Figure 3. The overview of the proposed joint network architecture. Given a composite image and a provided foreground mask, we first pass the input through an encoder for learning feature representations. The encoder is then connected to two decoders, including a harmonization decoder for reconstructing the harmonized output and a scene parsing decoder to predict pixel-wise semantic labels. In order to use the learned semantics and improve harmonization results, we concatenate the feature maps from the scene parsing decoder to the harmonization decoder (denoted as dot-orange lines). In addition, we add skip links (denoted as blue-dot lines) between the encoder and decoders for retaining image details and textures. Note that, to keep the figure clean, we only depict the links for the harmonization decoder, while the scene parsing decoder has the same skip links connected to the encoder.

a scene) in the target image. Based on the predicted scene parsing labels within the segmented target region, we find a region in the reference image that shares the same labels as the target region. The composite image is then generated by the color transfer method mentioned above (Figure 2(a)).

Discussions. With the above-mentioned data acquisition methods on three datasets, we are able to collect large-scale and high-quality training and test pairs (see Table 1 for a summarization). This enables us to train an end-to-end CNN for image harmonization with several benefits. First, our data collection method ensures that the ground truth images are realistic, so the network can really capture the image realism and adjust the input image according to the learned representations.

Another merit of our method is to enable quantitative evaluations. This is, we can use the synthesized composite image to measure errors by comparing to the ground truth images. Although there should be no single best solution for the image harmonization task, this quantitative measurement can give us a sense of how closer the images generated by different methods are, to a truly realistic image (discussed in Section 4), which is not addressed by previous approaches.

3.2. Context-aware Encoder-decoder

Motivated by the potential of the Context Encoders [20], our CNN learns feature representations of input images via an encoder and reconstruct the harmonized output results through a decoder. While the proposed deep network bears some resemblance, we add novel components for image harmonization. In the following, we present the objective function and proposed network architecture with discussion of novel components.

Objective Function. Given a RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and a provided binary mask $M \in \mathbb{R}^{H\times W\times 1}$ of the composite foreground region, we form the input $X \in \mathbb{R}^{H\times W\times 4}$ by concatenating I and M, where H and W are image dimensions. Our objective is to predict an output image $\hat{Y} = \mathcal{F}(X)$ that optimizes the reconstruction ($L2$) loss with respect to the ground truth image Y:

$$\mathcal{L}_{rec}(X) = \frac{1}{2}\sum_{h,w} \parallel Y_{h,w} - \hat{Y}_{h,w} \parallel_2^2 . \qquad (1)$$

Since the $L2$ loss is optimized with the mean of the data distribution, the results are often blurry and thus miss important details and textures from the input image. To overcome these problems, we show that adding skip links from the encoder to the decoder can recover those image details in the proposed network.

Network Architecture. Figure 3 shows basic components of our network architecture with an encoder and a harmonization decoder. The encoder is a series of convolutional layers and a fully connected layer to learn feature representations from low-level image details to high-level context information. Note that as we do not have any pooling layers, fine details are preserved in the encoder [20]. The decoder is a series of deconvolutional layers which aim to reconstruct the image via up-sampling from the representations learned in the encoder and simultaneously adjust the appearances of the foreground region.

However, image details and textures may be lost during the compression process in the encoder, and thus there is less information to reconstruct the contents of the input image. To retain those details, it is crucial that we add a skip link from each convolutional layer in the encoder to each corresponding deconvolutional layer in the decoder. We show this method is effectively useful without adding additional burdens for training the network. Furthermore, it can alleviate the problem of the $L2$ loss that prefers a blurry image solution.

Implementation Details. We implement the proposed network in Caffe [9] and use the stochastic gradient descent solver for optimization with a fixed learning rate 10^{-8}. In addition, we compute the loss on the entire image rather than the foreground mask to account for the reconstruction differences in the background region. We also try a weighted loss that considers the foreground region more importantly, but the results are similar and thus we use a simple loss function. Since the entire network is trained from scratch, we use the batch normalization [8] followed by a scaling layer and an ELU layer [5] after each convolutional and deconvolutional layers to facilitate the training process.

Discussions. We conduct experiments using the proposed network architecture with different input sizes. Interestingly, we find that the one with larger input size performs better in practice, and thus we use input resolution of 512×512. This observation also matches our intuition when designing the encoder-decoder architecture with skip links, where the network can learn more context information and details from a larger input image. To generate higher resolution results, we can up-sample the output of the network with joint bilateral filtering [22], in which the input composite image is used as the guidance to keep clear details and sharp textures.

3.3. Joint Training with Semantics

In the previous section, we propose an encoder-decoder network architecture for image harmonization. In order to further improve harmonization results, it is natural to consider semantics of the composite foreground region. The ensuing question is how to incorporate such semantics in our CNN, so that the entire network is still end-to-end trainable. In this section, we propose a modified network that can jointly train the image harmonization and scene parsing tasks simultaneously, while propagating semantics to improve harmonization results. The overall architecture is depicted in Figure 3, which adds the scene parsing decoder branch.

Joint Loss. In addition to the reconstruction loss described for image harmonization in (1), we introduce a pixel-wise cross-entropy loss with the standard softmax function $\mathbb{E}$ for scene parsing:

$$\mathcal{L}_{cro}(X) = -\sum_{h,w} \log(\mathbb{E}(X_{h,w};\theta)). \tag{2}$$

We then define a combined loss for both tasks and optimize it jointly:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{cro}, \tag{3}$$

where λ_i is the weight to control the balance between losses for image harmonization and scene parsing.

Network Architecture. We design the joint network by inheriting the encoder-decoder architecture described in the previous section. Specifically, we add a decoder to predict scene parsing results, while the encoder is to learn feature representations and is shared for both decoders. To extract semantic knowledge from the scene parsing model and help harmonization process, we concatenate feature maps from each deconvolutional layer of the scene parsing decoder to the harmonization decoder, except for the last layer which focuses on image reconstruction. In addition, skips links [19] are also connected to the scene parsing decoder to gain more information from the encoder.

Implementation Details. To enable the training process for the proposed joint network, both the ground truth images for harmonization and scene parsing are required. We then use a subset of the ADE20K dataset [31], which contains 12080 training images with the top 25 frequent labels. Similarly, training pairs for harmonization are obtained in a way described in the data acquisition section via color transfer.

To train the joint network, we start with the training data from the ADE20K dataset to obtain an initial solution for both the harmonization and scene parsing by optimizing (3). We set $\lambda_1 = 1$ and $\lambda_2 = 100$ with a fixed learning rate 10^{-8}. Next, we fix the scene parsing decoder with $\lambda_2 = 0$ and finetune rest of the network using all the training data introduced in Section 3.1 to achieve the optimal solution for image harmonization. Note that, during this fintuning step, the scene parsing decoder is able to propagate learned semantic information through the links between two decoders.

Discussions. With the incorporated scene parsing model, our network can learn the color distribution of certain semantic categories, e.g., the skin color on human or the sky-like colors. In addition, the learned background semantics can help identify which region to match for better foreground adjustment. During harmonization, it essentially uses these learned semantic priors to improve the realism of output results. Moreover, the incorporation of semantic information through joint training not only helps our image harmonization task, but also can be adopted to benefit other image editing tasks [30, 20].

Figure 4. Example results on synthesized datasets for the input, ground truth, three state-of-the-art methods and our proposed network. From the first row to the third one, we show one example for the MSCOCO, MIT-Adobe and Flickr datasets. Each result is associated with a PSNR score. Among all the methods, our harmonization results obtain the highest score.

Table 2. Comparisons of methods with mean-squared errors (MSE) on three synthesized datasets.

	MSCOCO	MIT-Adobe	Flickr
cut-and-paste	400.5	552.5	701.6
Lalonde [13]	667.0	1207.8	2371.0
Xue [28]	351.6	568.3	785.1
Zhu [32]	322.2	360.3	475.9
Ours (w/o semantics)	80.5	168.8	491.7
Ours	**76.1**	**142.8**	**406.8**

Table 3. Comparisons of methods with PSNR scores on three synthesized datasets.

	MSCOCO	MIT-Adobe	Flickr
cut-and-paste	26.3	23.9	25.9
Lalonde [13]	22.7	21.1	18.9
Xue [28]	26.9	24.6	25.0
Zhu [32]	26.9	25.8	25.4
Ours (w/o semantics)	32.2	27.5	27.2
Ours	**32.9**	**28.7**	**27.4**

To validate our scene parsing model, we compare the proposed joint network to a deeplab model [4], MSc-COCO-LargeFOV, that has a similar model capacity and size to our model but is initialized from a pre-trained model for semantic segmentation. We evaluate the scene parsing results on the validation set of the ADE20K dataset with the top 25 frequent labels. The mean intersection-over-union (IoU) accuracy of our joint network is 32.2, while the MSc-COCO-LargeFOV model achieves IoU as 36.0. Although our model is not specifically designed for scene parsing and is learned from scratch, it shows that our method performs competitively against a state-of-the-art model for semantic segmentation.

4. Experimental Results

We present the main results on image harmonization with comparisons to the state-of-the-art methods in this section. More results and analysis can be found in the supplementary material. The code, model and test set are available at https://github.com/wasidennis/DeepHarmonization.

Synthesized Data. We first evaluate the proposed method on our synthesized dataset for quantitative comparisons. Table 2 and 3 show the results of mean-squared errors (MSE) and PSNR scores between the ground truth and harmonized

Figure 5. Example results to show the comparison of our network with or without incorporating semantic information. With semantics, our result can recover the skin color and obtain higher PSNR score.

image. Note that it is the first quantitative evaluation on image harmonization, which reflects how close different results are to realistic images. We show that our joint network consistently achieves better performance compared to the single network without combining scene parsing decoder and other state-of-the-art algorithms [13, 28, 32] on all three synthesized datasets in terms of MSE and PSNR. In addition, it is also worth noticing that our baseline network without semantics already outperforms other existing methods.

In Figure 4, we show visual comparisons with respect to PSNR of the harmonization results generated from different methods. Overall, the harmonized images by the proposed methods are more realistic and closer to the ground truth images, with higher PSNR values. In addition, Figure 5 presents one comparison of our networks with and without incorporating the scene parsing decoder. With semantic understandings, our joint network is able to harmonize foreground regions according to their semantics and produce realistic appearance adjustments, while the one without semantics may generate unsatisfactory results in some cases.

Real Composite Images. To evaluate the effectiveness of the proposed joint network in real scenarios, we create a test set of 52 real composite images and combine 48 examples from Xue et al. [28], resulting in a total of 100 high-quality composite images. To cover a variety of real examples, we create composite images including various scenes and stylized images, where the composite foreground region can be an object or a scene.

We follow the same procedure as [28, 32] to set up a user study on Amazon Mechanical Turk, in which each user sees two randomly selected results at a time and is asked to choose the one that looks more realistic. For sanity checks, we use ground truth images from the synthesized dataset and heavily edited images to create easily distinguishable pairs that are used to filter out bad users. As a result, a total of 225 subjects participate in this study with a total of 10773 pairwise results (10.8 results for each pair of different methods on average). After obtaining all the pairwise results, we use the Bradley-Terry model (B-T model) [2, 12] to calculate the global ranking score for each method.

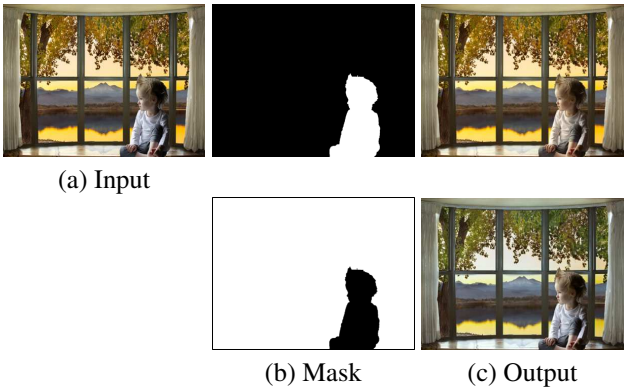


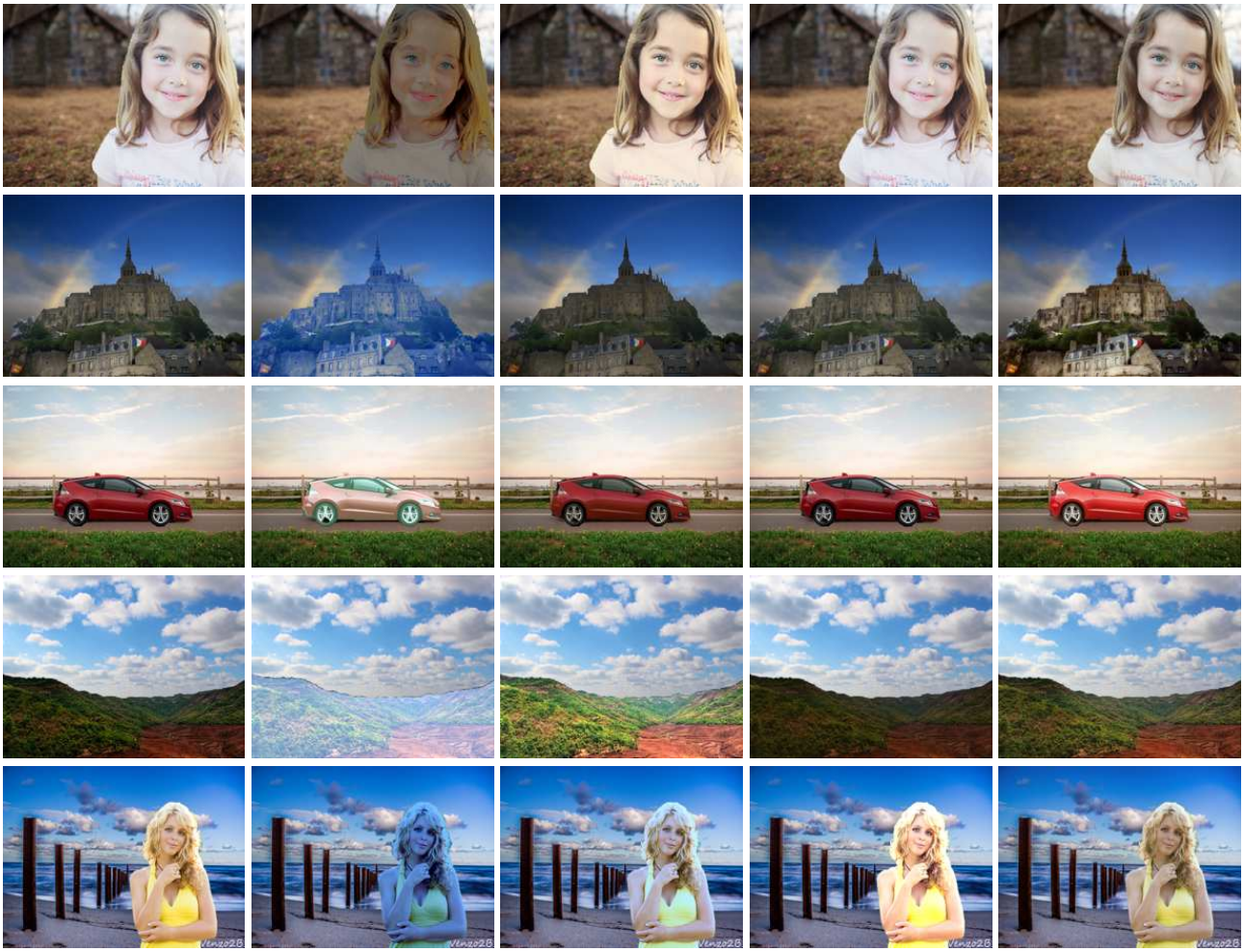
Figure 6. Given an input image (a), our network can adjust the foreground region according to the provided mask (b) and produce the output (c). In this example, we invert the mask from the one in the first row to the one in the second row, and generate harmonization results that account for different context and semantic information.

Table 4. Comparisons of methods with B-T scores on real composite datasets.

Dataset	[28]	Our test set	Overall
cut-and-paste	1.080	1.168	1.139
Lalonde [13]	0.557	0.067	0.297
Xue [28]	1.130	0.885	1.002
Zhu [32]	0.875	0.867	0.876
Ours	**1.237**	**1.568**	**1.424**

Table 4 shows that our method achieves the highest B-T score in terms of realism compared to state-of-the-art approaches on both our created test set and examples from [28]. Interestingly, our method is the only one that can improve the harmonization result with a significant margin from the input image (by cut-and-paste).

Figure 7 shows sample harmonized images by the evaluated methods. Overall, our joint network produces realistic output images, which validates the effectiveness of using synthesized data to directly learn how to harmonize composite images from realistic ground truth images. The results from [28] may be easily affected by the large appearance difference between the background and foreground regions during matching. For the method [32], it may generate unsatisfactory results due to the errors introduced during realism prediction, which may affect the color optimization step. In contrast, our network adopts a single feed-forward scheme learned from a well-constructed training set, and utilizes semantic information to improve harmonization results. The complete results on the real composite test set are presented in the supplementary material.



Input Lalonde [13] Xue [28] Zhu [32] Ours

Figure 7. Example results on real composite images for the input, three state-of-the-art methods and our proposed network. We show that our method produces realistic harmonized images by adjusting composite foreground regions containing various scenes or objects.

Generalization to Background Masks. With the provided foreground mask, our network can learn context and semantic information while transforming the composite image to a realistic output image. Therefore, our method can be applied to any foreground masks containing arbitrary objects, scenes or clutter backgrounds. Figure 6 illustrates one example, where originally the adjusted foreground region is the *child*. Instead, we can invert the mask and focus on harmonizing the region of *inverted child*. The result shows that our network can produce realistic outputs from different foreground masks.

Runtime Performance. Previous harmonization methods rely on matching statistics [13, 28] or optimizing an adjustment function [32], which usually require longer processing time (more than 10 seconds with a 3.4GHz Core Xeon CPU) on a 512×512 test image. In contrast, our proposed CNN is able to harmonize an image in 0.05 seconds with a Titan X GPU and 12GB memory, or 3 seconds with a CPU.

5. Concluding Remarks

In this paper, we present a novel network that can capture both the context and semantic information for image harmonization. We demonstrate that our joint network can be trained in an end-to-end manner, where the semantic decoder branch can effectively provide semantics to help harmonization. In addition, to facilitate the training process, we develop an efficient method to collect large-scale and high-quality training pairs. Experimental results show that our method performs favorably on both the synthesized datasets and real composite images against other state-of-the-art algorithms.

Acknowledgments. This work is supported in part by the NSF CAREER Grant #1149783, NSF IIS Grant #1152576, and a gift from Adobe. Portions of this work were performed while Y.-H. Tsai was an intern at Adobe Research.

References

[1] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *PAMI*, 35(8):1798 – 1828, 2013. 2

[2] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952. 7

[3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, 2011. 3

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 6

[5] D. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *ICLR*, 2016. 5

[6] S. J. Hwang, A. Kapoor, and S. B. Kang. Context-based automatic local image enhancement. In *ECCV*, 2012. 1

[7] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color!: Joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph. (proc. SIGGRAPH)*, 35(4), 2016. 2

[8] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 5

[9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5

[10] M. K. Johnson, K. Dale, S. Avidan, H. Pfister, W. T. Freeman, and W. Matusik. Cg2real: Improving the realism of computer generated images using a large collection of photographs. *IEEE Trans. Vis. Comp. Graph.*, 17(9), 2011. 2

[11] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016. 3

[12] W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang. A comparative study for single image blind deblurring. In *CVPR*, 2016. 7

[13] J.-F. Lalonde and A. A. Efros. Using color compatibility for assessing image realism. In *ICCV*, 2007. 1, 2, 6, 7, 8

[14] G. Larsson, M. Maire, and G. Shakhnarovich. Learning representations for automatic colorization. In *ECCV*, 2016. 2

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 3

[16] J.-Y. Lee, K. Sunkavalli, Z. Lin, X. Shens, and I. S. Kweon. Automatic content-aware color and tone stylization. In *CVPR*, 2016. 3

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 3

[18] S. Liu, J. Pan, and M.-H. Yang. Learning recursive filters for low-level vision via a hybrid neural network. In *ECCV*, 2016. 2

[19] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 5

[20] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders : Feature learning by inpainting. In *CVPR*, 2016. 1, 2, 3, 4, 5

[21] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Trans. Graph. (proc. SIGGRAPH)*, 22(3), 2003. 2

[22] G. Petschnigg, R. Szeliski, M. Agrawala, M. Cohen, H. Hoppe, and K. Toyama. Digital photography with flash and no-flash image pairs. *ACM Trans. Graph. (proc. SIGGRAPH)*, 23(3), 2004. 5

[23] F. Pitié and A. Kokaram. The linear monge-kantorovitch linear colour mapping for example-based colour transfer. In *CVMP*, 2007. 2

[24] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Comp. Graph. Appl.*, 21(5):34–41, 2001. 2

[25] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister. Multi-scale image harmonization. *ACM Trans. Graph. (proc. SIGGRAPH)*, 29(4), 2010. 1, 2

[26] M. W. Tao, M. K. Johnson, and S. Paris. Error-tolerant image compositing. *IJCV*, 103(2):178–189, 2013. 2

[27] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, and M.-H. Yang. Sky is not the limit: Semantic-aware sky replacement. *ACM Trans. Graph. (proc. SIGGRAPH)*, 35(4), 2016. 1, 2

[28] S. Xue, A. Agarwala, J. Dorsey, and H. Rushmeier. Understanding and improving the realism of image composites. *ACM Trans. Graph. (proc. SIGGRAPH)*, 31(4), 2012. 1, 2, 6, 7, 8

[29] Z. Yan, H. Zhang, B. Wang, S. Paris, and Y. Yu. Automatic photo adjustment using deep neural networks. *ACM Trans. Graph.*, 2015. 1

[30] R. Zhang, P. Isola, and A. A. Efros. Colorful image colorization. In *ECCV*, 2016. 2, 3, 5

[31] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ADE20K dataset. *CoRR*, abs/1608.05442, 2016. 3, 5

[32] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros. Learning a discriminative model for the perception of realism in composite images. In *ICCV*, 2015. 1, 2, 6, 7, 8