# Image Harmonization with Attention-based Deep Feature Modulation

Guoqing Hao
hao_guoqing@cvlab.cs.tsukuba.ac.jp
Satoshi Iizuka
iizuka@cs.tsukuba.ac.jp
Kazuhiro Fukui
kfukui@cs.tsukuba.ac.jp
University of Tsukuba, Japan

**Abstract**

We present a learning-based approach for image harmonization, which allows for adjusting the appearance of the foreground to make it compatible with background. We consider improving the realism by adjusting the high-level feature statistics of the foreground according to those of the background, which is motivated by the fact that specific image statistics between the foreground and background typically match in realistic composite images. Based on a fully convolutional network, we propose a novel attention-based module that aligns the standard deviation of the foreground features with that of the background features, capturing global dependencies in the entire image. This module is easily inserted into any types of convolutional neural networks, and allows improving the harmony of the composites with only a small additional computational cost. Experimental results on the image harmonization dataset and real composite images show that our method outperforms existing methods both quantitatively and qualitatively. Furthermore, in our experiment, our module is able to boost existing harmonization networks by simply inserting it into intermediate layers of those networks.

## 1 Introduction

Image composition is a commonly used operation in photo editing, where the composite image is generated by pasting a foreground region of a source image onto the background of a target image. However, generating composite images without adjusting their appearance often makes the composite image unrealistic. Therefore, image harmonization that makes the foreground compatible with the background is an important technique to generate realistic composites.

Although various approaches have been proposed for image harmonization [3, 4, 12, 23, 29, 32], it remains a challenging task as it requires high-level scene understanding for the composite realism. In this work, we consider improving the realism by adjusting the feature statistics of the foreground according to those of the background (Fig. 1). This is motivated by the fact that specific image statistics between the foreground and background typically match in a realistic composite [12, 29]. Unlike the existing methods that use low-level

Input DoveNet [3] Ours Ground Truth

Figure 1: Our method can adjust the appearances of foreground region to make it compatible with the background region. Given an input composite image, our method can generate harmonized image closer to the ground truth image.

feature statistics such as a histogram of luminance, we leverage the statistics of learned high-level features obtained by a convolutional neural network, which is able to exploit semantic information for the harmonization. Additionally, in this process, it is important to assess the strength of the relationship between the foreground and background regions. For example, a red car in the background has a small dependency with a woman dressed in yellow clothes in the foreground. We need to pay more attention to the feature statistics of the background contents that are similar to the foreground contents.

Based on the observations, we propose an effective way to adjust the feature statistics in a learning-based framework. We leverage a fully convolutional network as the basis of our approach, and propose a novel attention-based module that modulates the statistics of intermediate feature maps of the network based on the global dependencies between the foreground and background features. In this module, non-local similarity across image regions are first calculated using self-attention [31], which allows the network to incorporate non-local information from the entire composite image. Afterwards, the foreground and background regions are separated by using a corresponding mask, and then the standard deviation of the foreground features is modulated according to that of the similarity-weighted background features.

Our modulation layer has three advantages: First, it is able to capture global dependencies between the foreground and background, which is necessary to assess if the composite is realistic as a whole. Second, as the module is differentiable, the entire network including the module is able to be trained in an end-to-end fashion. Finally, this module can be easily inserted into any type of convolutional neural networks with only a small additional computational cost. In our experiment, our module boosts existing harmonization networks [2, 23] by simply inserting it into intermediate layers of those networks. To the best of our knowledge, it is the first attention-based feature modulation technique introduced to image harmonization. We conducted comprehensive experiments on the image harmonization dataset, and find that our method outperforms the existing methods via criteria of peak signal-to-noise ratio (PSNR). We also show results on real composite images. Since there are no available ground truth images for real composite images, we compare our model against existing methods with a perceptual user study, where our method shows better performance than existing methods.

# 2 Related Work

## 2.1 Image Harmonization

The problem of image harmonization is to generate composites that have a good match in the appearance between foreground and background regions. Some early approaches focused on matching low-level appearance statistics to generate realistic composite images, such as applying gradient domain methods [19], matching multi-scale statistics [22]. Lalonde and Efros [12] predicted the visual realism of composite images by learning color statistics from natural images and shifting the color statistics of foreground appearances to improve chromatic compatibility. Xue *et al*. [29] performed human subject experiments to identify the most significant statistics that are used to adjust foreground appearances accordingly. Zhu *et al*. [32] learned a deep convolutional neural network model [21] to predict realism score assessments of composite images and incorporated the realism score into an optimization function for improving the realism of composite images.

Recently, Tsai *et al*. [23] proposed an end-to-end neural network of image harmonization, in which a segmentation branch is incorporated into an end-to-end harmonization network to use semantic information. To learn the differences between various low-level features in the composite images, Cun and Pun [4] proposed a novel attention module to learn the attended foreground and background features separately. Cong *et al*. [3] released a large image harmonization dataset (iHarmony4) and introduced a novel domain verification discriminator for adversarial training. Wu *et al*. [27] leveraged the strengths of classical gradient-based approach [2] and generative adversarial network [5] (GAN) to solve the problem of high-resolution image blending. More recently, [6] and [13] introduced video harmonization task which intends to improve the realism of a composite video by performing appearance adjustments on the foreground. Different from existing methods, we consider improving the harmony by adjusting the high-level feature statistics of the foreground according to those of the background.

## 2.2 Attention

Similar to human attention, attention mechanisms for neural networks allow the model to focus on the most informative parts of the input according to the analysis of the features. Parmar *et al*. [17] introduced self-attention for image generation, in which pixel locations are explicitly encoded. Zhang *et al*. [31] simplified this method to not avoid explicitly encoding the pixel locations. Wang *et al*. [26] computed the similarity of objects in the different video frames with a self-attention mechanism. To learn the specific region in spatial space, Liu *et al*. [14] proposed partial convolution for learning masked region to image inpainting. Cun and Pun [4] proposed $S^2$AM to learn the difference between foreground and background in image harmonization. More related to our approach is source-reference attention [8] for video colorization, in which the similarity between the input source video frames and reference images is computed. We follow the same concept as existing attention methods and extend it to calculate the similarity between foreground and background regions.

## 2.3 Feature Map Modulation

Over the past decade, learning-based approaches have made massive progress in advancing the state of the art in a variety of domains. However, training a deep network from scratch

is significantly time-consuming. Ioffe and Szegedy [9] introduced a batch normalization layer that significantly eases the training of deep networks by normalizing feature statistics. Since then, several alternative normalization schemes [1, 24, 25, 28] have been proposed to improve its effectiveness.

Huang and Belongie [7] proposed an extension to normalization called adaptive instance normalization (AdaIN). AdaIN layer performs style transfer in the feature space by modulating intermediate feature statistics. Specifically, it scales the normalized content input with the standard deviation of style input, and shifts it with the mean of style input. Karras *et al*. [10] incorporated the AdaIN layer into a generator network so that the generator can lead to automatic, unsupervised separation of high-level attributes from stochastic variation in the generated images. They further improved their method by abandoning the normalization part of AdaIN to remove normalization artifacts [11]. Park *et al*. [16] introduced a novel generator based on AdaIN architecture for semantic image synthesis. Inspired by existing works on this topic, we further incorporate feature map modulation into image harmonization, in which the feature map modulation layer is employed to align the standard deviation of the foreground features with that of the background features. To the best of our knowledge, it is the first work that explores the feature map modulation in image harmonization.

# 3 Approach

Our learning-based framework is formed exclusively by convolutional layers. An overview of our network architecture is illustrated in Fig. 2. In detail, the encoder-decoder structure with skip connections [20] that concatenate all channels of the feature maps is taken as the backbone of our network. Afterwards, by interpolating the proposed module, the entire network can learn the appearance adjustments of the foreground and background. In this section, we first introduce the full specification of the backbone network. Then, we discuss the attention-based foreground-background feature map modulation layer.

## 3.1 Model

Our model is based on fully convolutional networks [15] that can be applied to images of any resolution. The input is a composite image with a binary channel that indicates the foreground masks (in which 1 and 0 indicate the region of foreground and background, respectively). Our network decreases the resolution to $1/8$ of the original size in three steps. Afterwards, with the interpolation of the proposed module, the feature statistics of the foreground is modulated according to those of the similarity-weighted background. Then, the output is restored to the original resolution in another three steps. Down-sampling is done by using convolutions with $2 \times 2$ pixel strides, while up-sampling is done by using the combination of the bilinear up-sampling and the convolutional layer. All convolutional layers use $3 \times 3$ kernel convolutions. Each convolution layer except for the last layer is followed by batch normalization [9] and Rectified Linear Unit (ReLU), while the last one uses a hyperbolic tangent activation function to keep the output in the [-1, 1] range. Finally, the input is added to the output of the network and clamped to be in the [0, 1] range.

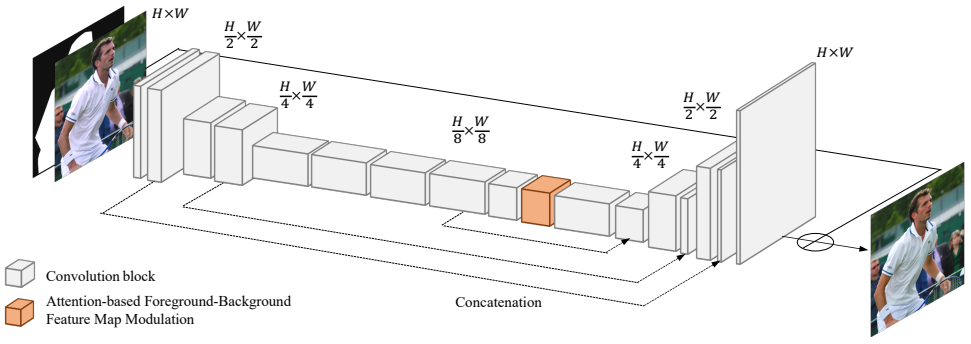



Figure 2: **An overview of our network architecture.** The model takes a composite image and mask as input. Attention-based foreground-background feature map modulation layer is used to perform modulation at the feature space. This layer allows modulating the feature map of foreground according to the similarity-weighted background. Finally, the output of the network is a harmonized image.

## 3.2 Attention-based Foreground-background Feature Map Modulation

We discuss the details of the attention-based foreground-background feature map modulation layer in this section. In the image harmonization task, non-local information across image regions plays a critical role because the harmonization requires the use of non-local features from the entire image. Increasing the size of the convolution kernel enables the calculation of non-local information across image regions. However, this increase leads to a loss in computational and statistical efficiency. Instead of increasing the size of convolution kernels, we utilize self-attention [31] to calculate the non-local information from the entire image.

In order to learn the features of foreground and background individually, we separate the foreground and background with the corresponding mask. After that, to assess the strength of the relationship between the foreground and background regions, we extend the attention mechanism [8, 31] to be able to use non-local similarities between the foreground and background regions.

Finally, inspired by existing work [7, 10, 11] on feature map modulation, we introduce the feature map modulation (FMM) layer to align the standard deviation of the foreground features with that of the similarity-weighted background features.

More formally, let the feature map produced by the self-attention block be $h \in \mathbb{R}^{C\times H\times W}$ with $C$ channels, height $H$ and width $W$, and let the binary mask be $M \in \mathbb{R}^{1\times H\times W}$ with 1 channel, height $H$ and width $W$. We first separate foreground and background with the corresponding mask:

$$\begin{aligned} h_f &= h \times M \\ h_b &= h \times (1-M), \end{aligned} \tag{1}$$

where the $h_f$ and $h_b$ are the feature maps of the foreground and background, respectively. Then, to pay more attention to the feature statistics of background contents that are similar to the foreground contents, the similarity-weighted feature map of background $h_b^a$ can be

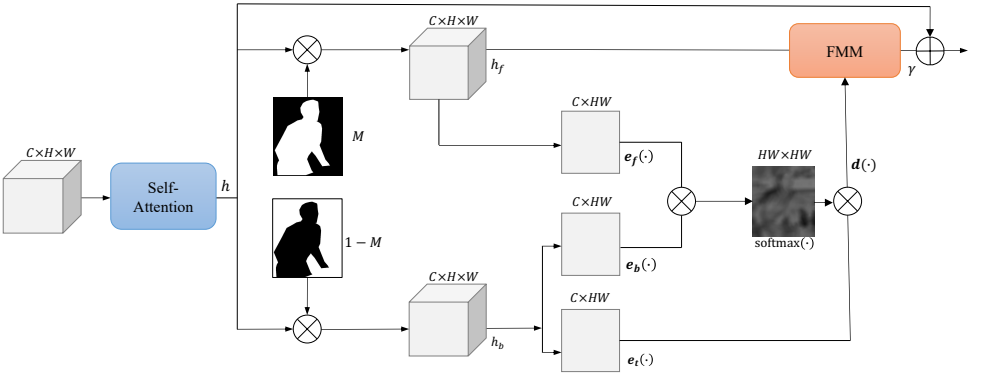



Figure 3: **An overview of the attention-based foreground-background feature map modulation module.** This module first utilizes a self-attention block [31] to calculate non-local information from the entire image. Afterwards, the foreground and background regions are separated by using a corresponding mask, and then the standard deviation of the foreground features is modulated according to that of the similarity-weighted background features. The modulation is done with the feature map modulation (FMM).

defined as:

$$h_b^a = d\left(e_t\left(h_b\right)\text{softmax}\left(e_b\left(h_b\right)^\top e_f\left(h_f\right)\right)\right), \tag{2}$$

where $e_t, e_b, e_f$ are encoding functions that use a convolution operator with $1 \times 1$ kernels, and $d : \mathbb{R}^{C\times HW} \rightarrow \mathbb{R}^{C\times H\times W}$ is a decoding function that reshapes the tensor. After that, we employ the feature map modulation layer to modulate the feature statistics of the foreground according to those of the similarity-weighted background:

$$FMM(h, h_f, h_b^a) = \gamma \times (h_f \times \frac{\sigma_{h_b^a}}{\sigma_{h_f}}) + h, \tag{3}$$

where $\gamma \in \mathbb{R}$ is a learned parameter, $\sigma$ indicates the standard deviation computed across spatial dimensions independently for each channel and each sample, $h$ is the feature map produced by the self-attention block. We add a skip connection between the outputs of self-attention and FMM because learning only residuals of the module can be easy to converge. A visual overview of the attention-based foreground-background feature map modulation layer is shown in Fig. 3.

Intuitively, we consider that the attention-based foreground-background feature map modulation layer detects the appearance style of the similarity-weighted background regions and apply the detected appearance style to the foreground. Armed with attention-based foreground-background feature map modulation layer, the network is able to improve the harmony of the composite images.

## 3.3 Optimization

Previous methods [3, 4] rely on adversarial training frameworks to learn image harmonization. However, one of the main issues of the adversarial training framework is the instability during learning, which leads to loss of generalization. To avoid this issue, we train the model

exclusively with MSE loss as the criterion in the pixel domain. The MSE loss used for training is described as following:

$$\arg\min_{\theta} \mathrm{E}\,\|M(x;\theta)-y\|^2,$$

where $\theta$ is the parameter of the image harmonization network $M$. $x$ and $y$ are the input and ground truth images, respectively.

Optimization is done by using the ADADELTA[30] algorithm, which is a variant of stochastic gradient descent that heuristically estimates the learning rate parameter, thus requiring no hyper-parameters to tune.

Table 1: Comparisons between our method against existing methods, on the iHarmony4 dataset[3], based on criteria of PSNR.

| Method | HCOCO | HAdobe5k | HFlickr | Hday2night | All |
|---|---|---|---|---|---|
| Lalonde and Efros [12] | 31.14 | 29.66 | 26.43 | 29.80 | 30.16 |
| Xue *et al.* [29] | 33.32 | 28.79 | 28.32 | 31.24 | 31.40 |
| Zhu *et al.* [32] | 33.04 | 27.26 | 27.52 | 32.32 | 30.72 |
| DIH [23] | 34.69 | 32.28 | 29.55 | 34.62 | 33.41 |
| $S^2$AM [4] | 35.47 | 33.77 | 30.03 | 34.50 | 34.35 |
| DoveNet [3] | 35.83 | 34.34 | 30.21 | **35.18** | 34.75 |
| Ours | **36.87** | **34.99** | **33.36** | 34.31 | **35.86** |

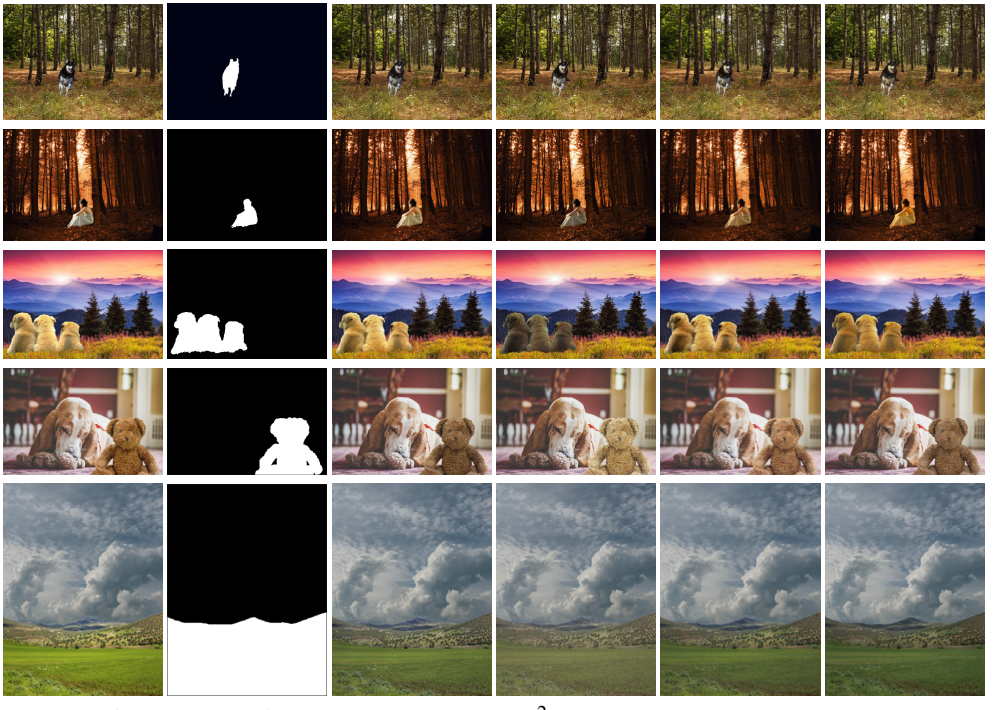
# 4 Results

Our model is implemented with PyTorch [18]. We train our model on the iHarmony4 dataset[3] that contains four subsets: HCOCO, HAdobe5k, HFlickr, Hday2night. In total, this dataset contains 65742 images for training and 7404 images for testing. Following the previous work [3], we merge the training sets of all four sub-datasets as one whole dataset for training, while the evaluation is performed on the test set of each sub-dataset and the whole test set. The entire network is trained by using patches of $256 \times 256$ pixels and a batch size of 24 for 630,000 iterations. We use average scores of PSNR over the test set as an evaluation metric, and compute the PSNR based on $256 \times 256$ images, following the existing method [3]. We omitted the mean square error (MSE) metric from here, as the PSNR is calculated based on MSE. Comparisons based on MSE can be found in the supplementary material.

Table 2: Our proposed layer works with different baselines. The methods with "+" represent its respective method with our proposed module. The "baseline" represents our backbone network. (our full method without the proposed module.)

| Method | DIH [23] | DIH+ | $S^2$AM [4] | $S^2$AM+ | baseline | baseline+ |
|---|---|---|---|---|---|---|
| PSNR↑ | 33.02 | 34.13 | 34.31 | 35.26 | 32.98 | **35.86** |

**Quantitative Evaluation.** We evaluate our method on the iHarmony4 dataset. The quantitative results based on PSNR are shown in Table 1. Notice that, all results except ours



Composite Mask DIH [23] $S^2$AM [4] DoveNet [8] Ours

Figure 4: Our method can adjust the appearances of foreground region to make it compatible with the background region. Given the real composite image and corresponding mask, our method can generate more realistic image than existing methods.

are taken from DoveNet [8]. From which, we can observe that our method outperforms the existing methods in most cases.

**Qualitative Analyses.** In Fig. 4 and Fig. 5, we show qualitative comparisons of the competing methods on the iHarmony4 test set and real composite images. We find that our method generates composite images with much better visual realism both on the test set and real composite images. More results on the test set and real composite images can be found in the supplementary material.

Table 3: Numerical comparisons of methods with perceptual user study on real composite images.

| Methods | Lalonde and Efros [12] | DIH [23] | $S^2$AM [4] | DoveNet [8] | Ours |
|---|---|---|---|---|---|
| Total votes | 80 | 147 | 208 | 238 | **317** |
| Average votes | 8% | 15% | 21% | 24% | **32%** |

**Effectiveness of the proposed module.** Our proposed module is simply inserted into convolutional neural networks with only a small additional computational cost. To quantify the importance of the attention-based foreground-background feature map modulation layer, we conduct experiments by inserting the proposed layer into existing learning-based image harmonization networks: DIH [23] and $S^2$AM [4]. DIH and $S^2$AM base on encoder-decoder

Composite Mask DIH [23] $S^2$AM [9] DoveNet [3] Ours Ground Truth

Figure 5: Visual comparison of different methods on iHarmony4 [3], where the mask indicates the region to be composited.

architecture and GAN framework, respectively. Note that we train existing methods on the iHarmony4 [3] dataset with released implementation. As shown in Table. 2, the performance of the baselines enhanced with our proposed layer consistently outperforms its counterparts, in both encoder-decoder architecture and GAN framework.

Table 4: Ablation results of our proposed module on the iHarmony4 dataset [3]. The "baseline" stands for the backbone network in our full method. The "foreground-background FMM" stands for "remove self-attention block from our proposed module."

| Method | PSNR↑ |
|---|---|
| baseline | 32.98 |
| baseline + foreground-background FMM | 35.17 |
| baseline + self-attention | 35.06 |
| baseline + attention-based foreground-background FMM | **35.86** |

**Perceptual User Study.** We further compare our method with four existing methods on real composite images which are collected by [23, 29]. As there are no ground truth images available, we performed a user study to evaluate the performance on real composite images. Following the previous work [9], we used all 99 images for evaluation and processed them with 5 different approaches. Note that all learning-based approaches were trained on the iHarmony4 dataset. Afterwards, for each real composite image, we obtained 7 outputs, including the original copy-and-paste, corresponding mask, and the harmonized images of 5 methods. We invited 10 subjects to participate in this study, given the original copy-and-

Composite Mask Baseline + SA + F-b FMM Ours Ground Truth

Figure 6: Visual comparisons of each part in the proposed module. The '+ SA' and '+ F-b FMM' stand for 'baseline + self-attention' and 'baseline + foreground-background FMM', respectively.

paste image and corresponding mask, the subjects are told to pick the most realistic image among the harmonizaed images. Results are shown in Table. 3, we can observe that our mothod outperforms all existing methods.

**Ablation Study.** To evaluate the importance of each part in the attention-based foreground-background FMM, we conduct a ablation study on the inner structure of the proposed module. All experiments are tested on iHarmony4 [3] with the same configuration. As shown in Table.4, our full method outperforms other architectures. Besides numerical comparisons, we also show visual comparisons of each part in the proposed module for better understanding of the module. It is obvious from Fig. 6, that the baselines enhanced with each sub-module can generate images which are more harmonious than images generated by the baseline, and our full method can generate the most realistic harmonized images.

## 5 Conclusion

We proposed a novel attention-based foreground-background feature map modulation layer to improve the realism of composite images by adjusting the high-level feature statistics of the foreground according to those of the background. Experimental results showed that our proposed method outperforms the existing methods both quantitatively and qualitatively.

## References

[1] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint*, arXiv:1607.06450, 2016.

[2] Peter Burt and Edward Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on communications*, 31(4):532–540, 1983.

[3] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] X. Cun and C. Pun. Improving the harmony of the composite image by spatial-seperated attention module. *IEEE Transactions on Image Processing*, 29:4759–4771, 2020.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, 2014.

[6] Hao-Zhi Huang, Sen-Zhe Xu, Jun-Xiong Cai, Wei Liu, and Shi-Min Hu. Temporally coherent video harmonization using adversarial networks. *IEEE Transactions on Image Processing*, 29:214–224, 2019.

[7] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[8] Satoshi Iizuka and Edgar Simo-Serra. Deepremaster: Temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, 38(6):1–13, 2019.

[9] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[12] Jean-Francois Lalonde and Alexei A Efros. Using color compatibility for assessing image realism. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007.

[13] Donghoon Lee, Tomas Pfister, and Ming-Hsuan Yang. Inserting videos into videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[16] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[17] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2018.

[18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

[19] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 22(3):313–318, 2003.

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

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR*, 2015.

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

[23] 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 (CVPR)*, 2017.

[24] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint*, arXiv:1607.08022, 2016.

[25] Nanne van Noord and Eric Postma. A learned representation of artist-specific colourisation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (CVPRW)*, 2017.

[26] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[27] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. Gp-gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

[28] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

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

[30] Matthew D. Zeiler. ADADELTA: an adaptive learning rate method. *arXiv preprint*, arXiv:1212.5701, 2012.

[31] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2019.

[32] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A. Efros. Learning a discriminative model for the perception of realism in composite images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.