Photorealistic Neural Style Transfer: Generating Realistic Images without GANs

Richard R. Yang
Department of Computer Science
Stanford University
rry@stanford.edu

Abstract

This work introduces a new approach toward photorealistic style transfer. We observe that current style transfer techniques result in distortions and artifacts in the generated images. To address these issues, we propose a two-stage optimization process that transfers style globally and regionally. As evaluated by a user study, our results are generally comparable to the previous state-of-the-art method, but successfully handles their failure cases. We also use natural scene statistic metrics to quantitatively evaluate our results, and we outperform all previous methods under this metric. Code and additional results are available at https://github.com/richard-y/neural-photo-style

1 Introduction

Style transfer is the technique of transferring the style in a reference image to another image. Classic methods for style transfer involve algorithmic image transformations such as color transfer [13] and histogram matching [16], but are only applicable to specific instances with limited effectiveness. The work by Gatys et al. [2] shows that the correlation between features learned by a convolutional neural network (CNN) are effective in capturing the style and content of the reference photos, and inspired a new era of neural style transfer algorithms. While these neural algorithms are adept in producing new artistic renditions, we observe the images typically have distortions and artifacts that render them unrealistic.

In this work, we expand upon neural style transfer to tackle the challenge of photorealistic style transfer (PST). Our goal is to perform style transfer with the constraint that the result is visually realistic, as if it was captured by a camera in the real world. This technique is valuable to many real world applications, such as removing haze from a photo or converting a day-time photo to a night-time photo.

The contributions of this paper are:

1. We introduce a novel approach to PST by using a two-stage optimization process that transfers a global style and a regional style, and iteratively sharpening the generated image to constrain the search space of the optimization algorithm. Our approach is fully optimization-based and does not require pre-training on a specific dataset, thus it is generalizable to most images and scenarios.

2. We conduct a user study to evaluate our results against previous style transfer algorithms. We are also the first to quantitatively evaluate our results using natural scene statistics.

Preprint. Work in progress.

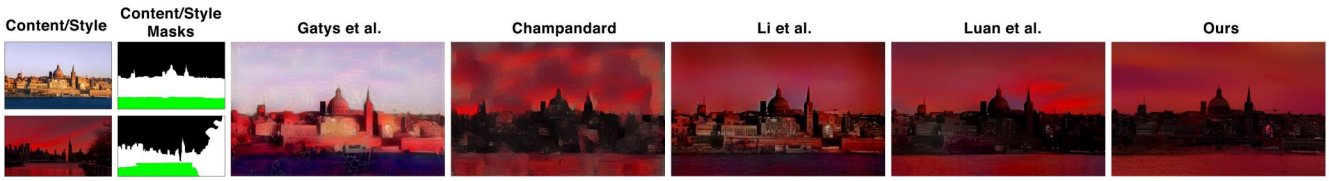

Figure 1: A comparison of results from current style transfer methods and ours with the using the same reference images.

2 Related Work

2.1 Optimization-based Neural Style Transfer

Gatys et al. introduce a neural style transfer (NST) technique, which shows that feature representations learned by a CNN encode both the style and content of an image [2]. This technique poses style transfer as an optimization problem, where we **generate** an image to minimize distance between its feature representations and that of the **style** and **content** reference images. This technique produces impressive artistic results, but the generated images often contain artifacts and distortions even if the reference images are photographs. Additionally, this technique suffers from a content mismatch problem where similar objects in the reference images may not be receive the same stylization. For example, consider reference images both with the sky and buildings in the scene. The generated image may have buildings in the content image with the style of the sky, e.g. column 3 of Figure 1.

Champandard expand upon the work of Gatys by utilizing semantic masks of the images in the optimization loss function to enforce style transfer within the same semantic regions [1]. This work reduces the content mismatch problem in NST, but the results still contain artifacts and distortions.

Li et al. is the first work toward photorealistic style transfer by incorporating an edge loss in the optimization loss function [8]. They perform edge detection on the images using the Laplacian operator, and minimize the difference between the edges in the generated and reference images. While this work is the first to reduce the distortions in generated images, the results are still very artistic.

2.2 Feed-forward Neural Style Transfer

As an alternative to the optimization-based style transfer techniques, Johnson et al. reformulate the NST problem and build an encoder-decoder network that learns to stylize an input image in one forward pass [6]. The results of this work allow for real-time style transfer during inference, but require multiple hours of training per style image.

The work by Luan et al. achieves state-of-the-art results in photorealistic style transfer. Their work expands upon the real-time style transfer from Johnson et al. by proposing a photorealistic regularization term that constrains the generated image to be represented through local affine color transformations of the content image with semantic masks [10]. This work is the first to significantly reduce the amount of artifacts and distortions in generated images, however, it suffers from various failure cases we show in the results section.

2.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are also adept in producing realistic images. GANs formulate image generation as a zero-sum game between a generator neural network and discriminator neural network [3]. Two specific architectures are applicable to the style transfer problem, Conditional Adversarial Networks [5] which requires pairs of images in two domains as input, and Cycle-Consistent Adversarial Networks [17] which does not require the explicit pairing. The effectiveness of these networks depend on the quality of the training data. In addition, the models expect the input image at evaluation time to come from the same domain as the training images, which is a limitation when using arbitrary photographs as input.

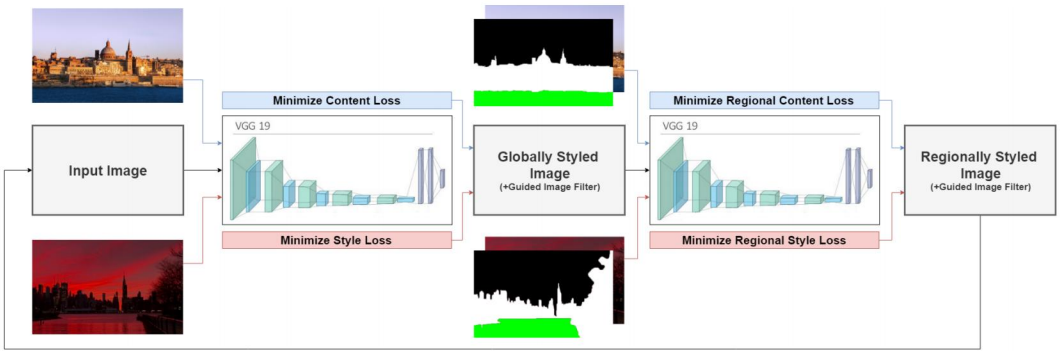

Figure 2: Our algorithm consists of two optimization stages: a global style transfer and regional style transfer. After each stage, we apply a post-processing smoothing filter to remove artifacts.

3 Methods

In our approach, we combine the intuition behind the techniques by Gatys et al. and Champandard. The weakness in the former is the content mismatch problem, while the weakness in the latter is the lack of a consistent style across the entire image. Therefore, we address both of these issues by having two stages in each optimization step as shown in Figure 2.

In the first stage, we perform a *global* style transfer based on the technique by Gatys et al. To transfer the style of a reference image *S* onto a content image *C* to produce an output image *O*, we minimize the following loss function [2]:

$$\mathcal{L}_{global} = \alpha \sum_{\ell=1}^{L} \mathcal{L}_{content}^{\ell} + \beta \sum_{\ell=1}^{L} \mathcal{L}_{style}^{\ell} \text{ where:}$$
$$\mathcal{L}_{content}^{\ell} = \frac{1}{2N_\ell D_\ell} \sum_{ij} (F_\ell[O] - F_\ell[C])_{ij}^2$$
$$\mathcal{L}_{style}^{\ell} = \frac{1}{2N_\ell^2} \sum_{ij} (G_\ell[O] - G_\ell[S])_{ij}^2$$

For L layers in a pre-trained CNN indexed by ℓ, $F_\ell[\cdot] \in \mathbb{R}^{N_\ell \times D_\ell}$ is a feature matrix with N_ℓ filters of D_ℓ vectorized feature maps of an input image. Then $G_\ell[\cdot] \in \mathbb{R}^{N_\ell \times N_\ell} = F_\ell[\cdot]F_\ell[\cdot]^T$ is the Gram matrix of the vectorized feature maps. The scalar weights α and β balance the contribution of the content and style losses to the total global loss. We use the VGG-19 network [14] as the pre-trained CNN for feature maps.

In the second stage, we perform a *regional* style transfer similar to the technique by Champandard [1]. We now introduce semantic maps of the content and style images as additional inputs, where each map has *R* distinct regions. For every pixel in the semantic map, we use k-means to cluster the pixel to one of the *R* semantic regions. After clustering, we generate *R* binary masking matrices indexed by region *r* and layer ℓ, $M_\ell^r[\cdot] \in \mathbb{R}^{N_\ell \times D_\ell}$, where the elements are 1 if the pixel is in region r or 0 otherwise. We apply the masking matrices to the feature matrices with element-wise multiplication, $F_\ell^r[\cdot] = F_\ell[\cdot] \odot M_\ell^r[\cdot]$. Subsequently, we also use the masked feature matrices for Gram matrix calculations. To sum up, our regional loss function becomes:

$$\mathcal{L}_{regional} = \alpha \sum_{\ell=1}^{L} \mathcal{L}_{content}^{\ell} + \beta \sum_{\ell=1}^{L} \mathcal{L}_{style}^{\ell} \text{ where:}$$
$$\mathcal{L}_{content}^{\ell} = \frac{1}{2N_\ell D_\ell} \sum_{r=1}^{R} \sum_{ij} (F_\ell^r[O] - F_\ell^r[C])_{ij}^2$$
$$\mathcal{L}_{style}^{\ell} = \frac{1}{2N_\ell^2} \sum_{r=1}^{R} \sum_{ij} (G_\ell^r[O] - G_\ell^r[S])_{ij}^2$$

For optimization, we use either the Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimizer [9] or the Adam optimizer [7] to generate the stylized images. After initial evaluation, we discover that there are often glitch-like artifacts in small sections of the image which diminish photorealism. This phenomenon is due to the local minima found by the optimizer, since we are

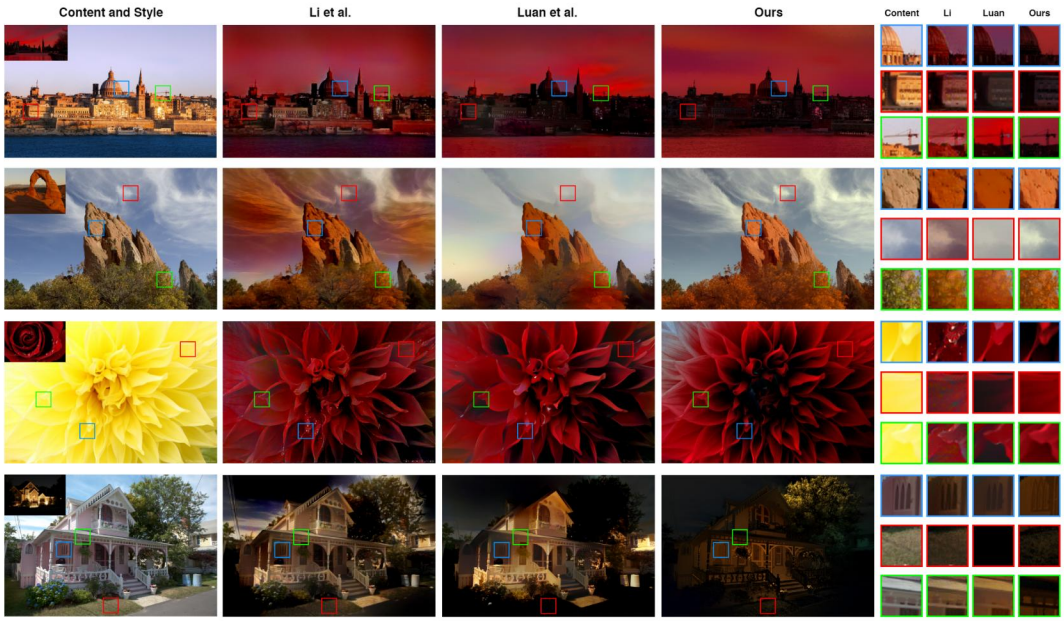

Figure 3: Comparison between our method and the two state-of-the-art photorealistic style transfer algorithms. Details are magnified at the colored boxes for comparison. Best viewed on a computer. Additional results are available at https://github.com/richard-y/neural-photo-style

changing the minimization objective after each stage. To address this issue, we apply a guided image filter after each stage to clean up any remnant artifacts. The guided image filter smooths a noisy image while preserving edges and structure from an input guidance image [4], which we choose as the original content image. We use the OpenCV implementation of this filter in our project.

3.1 Dataset

Since our approach is a fully optimization-based, there is no need for a training dataset. For evaluation, we use the same collection of images that Luan et al. use to evaluate their photorealistic style transfer algorithm [10]. This dataset contains 60 content and style image pairs, along with the corresponding semantic maps for each pair. The images are of varying dimensions and content, though most are photographs from the real world.

4 Results and Discussion

4.1 Qualitative Evaluation with Human User Study

To compare our stylized images with previous technique, we conduct a preliminary user study with 105 human respondents who voluntarily participated through a link shared on Facebook. We randomly select 5 content & style image pairs from the dataset, and generate stylized images using the style transfer techniques from [1] [2] [8] [10] and ours. We display the stylized images in random order and ask the respondent to *select one image that looks the most realistic, as if it was captured in the real world by a camera*. In figure 4, we show the results from the study. While slightly more users selected images from Luan et al. over ours in image sets (2) and (3), nearly all users selected ours in image set (5). Our images are comparable those of Luan et al. but our algorithm handles their failure cases well, such as image set (5). The exact images in set (5) are shown in row 2 of Figure 3.

4.2 Quantitative Evaluation with Natural Scene Statistics

In perception theory, natural images (i.e. images from the real world) tend to have specific statistical properties [11]. Previously, natural scene statistics have been used to evaluate the quality of image

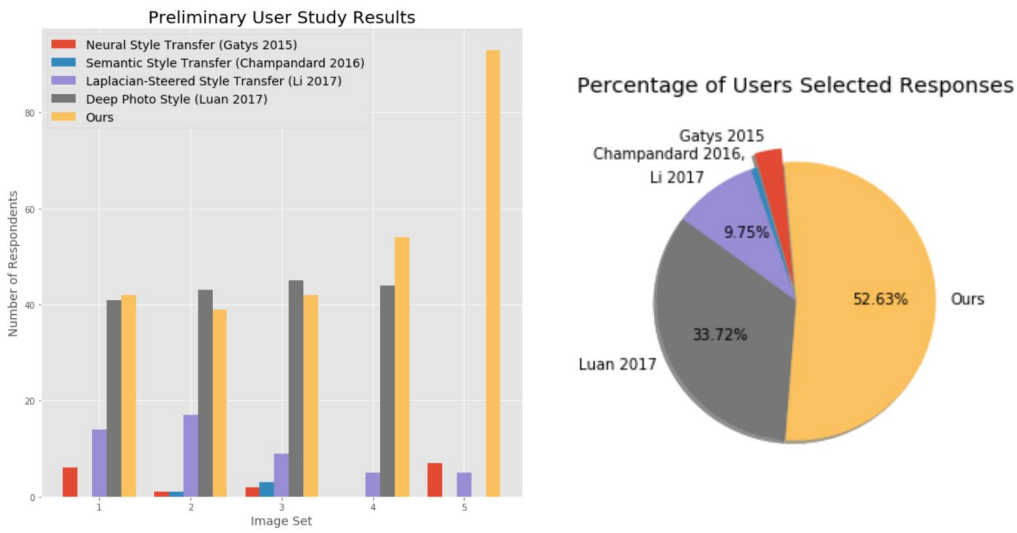

Figure 4: On the right, the number of users that selected images from each technique as the most *realistic* for each image pair. On the left, the percentage breakdown of all user responses.

Metric	Content Image	Gatys	Champandard	Li	Luan	Ours
Mean BRISQUE	19.806	29.625	30.433	23.211	21.121	**20.361**
Mean NIQE	2.325	4.236	4.230	3.417	3.169	**2.871**

Table 1: Mean BRISQUE and NIQE scores for all techniques and the original content image.

de-noising and compression algorithms [15]. Since the goal of this work is to generate photorealistic images, we use various natural scene statistical models as an image quality metric.

In the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), we train a support vector regression (SVR) model on a dataset of images with specific distortions (e.g. Gaussian noise, blurring) and a preset score for the severity of the distortion [11]. In the Natural Image Quality Evaluator (NIQE), we train another SVR model on a dataset of completely natural images without distortions [12]. BRISQUE returns a non-negative scalar score in the range [0,100], while NIQE returns a non-negative scalar score without upper bound. In both models, a lower value reflects a better perceptual quality of the image. We use the pre-trained BRISQUE and NIQE models from MATLAB to evaluate the images generated by each technique for all 60 image pairs in our evaluation dataset, and the mean scores are shown in Table 1. Our method achieves the lowest mean scores for both metrics in comparison to other methods.

4.3 Hyperparameters

The hyperparameters of our algorithm include the VGG-19 layers for style and content loss features, the style and content weights, and parameters r and ϵ for the guided image filter. We tune the style and content weights and the guided image filter parameters by performing a log-scale random search and selecting the set resulting in the lowest average NIQE score across three example images. We do not tune the VGG-19 layers and use the same as [2] since there are $\binom{19}{19} + \binom{19}{18} + ... + \binom{19}{2} + \binom{19}{1}$ possible subsets of layers to consider.

5 Conclusion

We introduce a novel approach to photorealistic style transfer using a two-stage optimization process. From a preliminary user study (which will be expanded upon in the future), our technique is comparable to the current state-of-the-art but handles their failure cases. Using NSS metrics as a quantitative evaluation, our algorithm is superior to all previous methods.

Contributions and Acknowledgements

This work was completed solely by Richard Yang for the final project of CS 230: Deep Learning. I would like to acknowledge the authors of [1], [8], and [10] for open-sourcing their code and models, which allowed the comparison between the different techniques to be possible. I also would like to acknowledge Steven Chen and Daniel Kunin for their insights, and the 105 users who participated in the user study to evaluate the results.

References

[1] Alex J Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016.

[2] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

[3] 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*, pages 2672–2680, 2014.

[4] Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE transactions on pattern analysis & machine intelligence*, (6):1397–1409, 2013.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

[6] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[8] Shaohua Li, Xinxing Xu, Liqiang Nie, and Tat-Seng Chua. Laplacian-steered neural style transfer. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1716–1724. ACM, 2017.

[9] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[10] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. *CoRR, abs/1703.07511*, 2, 2017.

[11] Anish Mittal, Anush K Moorthy, and Alan C Bovik. Blind/referenceless image spatial quality evaluator. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 723–727. IEEE, 2011.

[12] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a" completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013.

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

[14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[15] Garrett B Stanley, Fei F Li, and Yang Dan. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *Journal of Neuroscience*, 19(18):8036–8042, 1999.

[16] Hanli Zhao, Xiaogang Jin, Jianbing Shen, and Feifei Wei. Real-time photo style transfer. In *CAD/Graphics*, pages 140–145, 2009.

[17] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.