Multi-stage Optimization for Photorealistic Neural Style Transfer

Richard R. Yang
Department of Computer Science
Stanford University
richard.yang@cs.stanford.edu

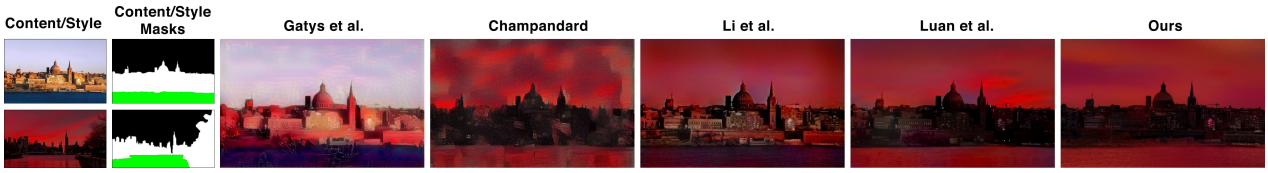

Figure 1: A survey of existing style transfer techniques applied to the task of photorealistic style transfer, where the generated image should seem as if it was taken in the real world.

Abstract

This work introduces a new approach toward photorealistic style transfer. When applying current style transfer techniques on real world photographs, the generated results often contain distortions and artifacts that diminish the real-world quality of the photograph. To address these issues, we propose a two-stage optimization process that transfers style globally and regionally and applies a sharpening filter after each step. As evaluated by a user study, our method is qualitatively comparable to existing state-of-the-art methods, but successfully handles previous failure cases. Our method also quantitatively outperform previous methods as evaluated by natural scene statistic metrics.

1. Introduction

Style transfer is the technique of transferring the style of a reference image to another image. Classic methods for style transfer involve algorithmic image transformations such as color transfer [18] and histogram matching [23], but are only applicable to specific instances with limited effectiveness. Gatys et al. [5] show that the correlation between features learned by a convolutional neural network (CNN) are effective at capturing the style and content of 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 that the images typically have distortions and artifacts, rendering 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 for many real world applications, such as removing haze or converting a day time photo to a night time photo.

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 training a style transfer network on a specific dataset or the use of generative adversarial networks (GANs), and is thus generalizable to most images and scenarios.

2. Related Work

2.1. Optimization-based Neural Style Transfer

Gatys et al. introduce a neural style transfer (NST) technique, which demonstrates that feature representations learned by a CNN encode both the style and content of an image [5]. 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 ob-

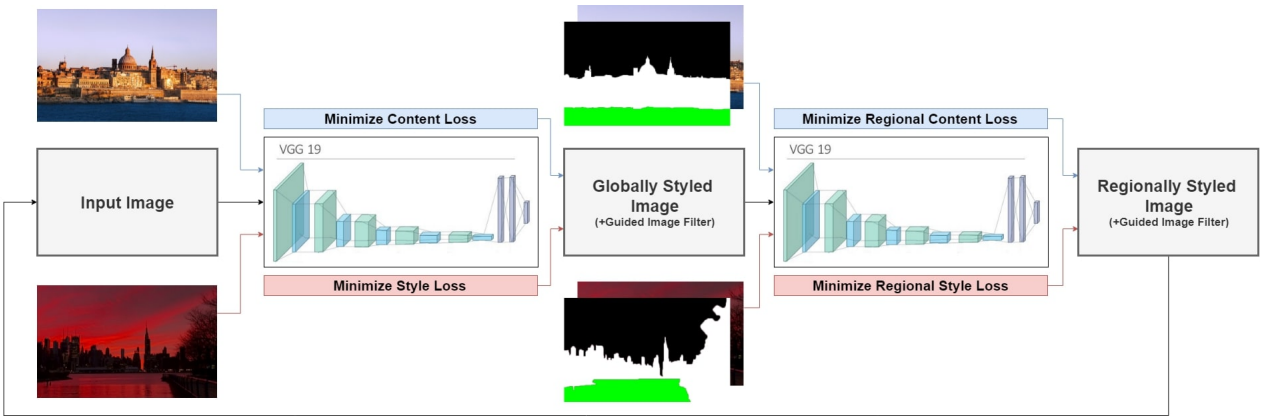

Figure 2: Our algorithm consists of two optimization stages: a global style transfer and regional style transfer. After each stage, we apply a post-processing smoothing filter to remove artifacts.

jects in the reference images may not receive the same stylization. For example, consider reference images with both 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. Figure 1 Column 3.

The subsequent work of Gatys et al. [6] and Champandard [2] expand upon NST by utilizing semantic masks of the images to enforce style transfer within the same semantic content regions. These works introduce masking matrices in in the optimization loss function to enforce spatial control of style transfer. While masking matrices are effective in reducing the content mismatch problem in NST, the results still contain artifacts and distortions that diminish realism.

Li et al. is the first work toward photorealistic style transfer by incorporating an edge loss in the optimization loss function [13]. They perform edge detection on the images using the Laplacian, and minimize the difference between edges in the generated and reference images. While this work is the first to reduce the distortions and preserve lines in generated images, the results can still be attributed as artistic and not photorealistic.

There are several variations of the techniques we previously described. We refer the reader to [10] for a full comprehensive review of related NST works.

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 [11]. This allows for real-time style transfer during inference, but requires multiple hours of training per style image.

Luan et al. achieves state-of-the-art results in photorealistic style transfer [15]. Their work expands upon the real-time style transfer from Johnson et al. by proposing a photorealistic regularization term that constrains the generated image through local affine color transformations of the content image with semantic masks [15]. This work is the first to significantly reduce the amount of artifacts and distortions in generated images and is the current state-of-the-art technique for photorealistic style transfer.

2.3. Generative Adversarial Networks

Generative Adversarial Networks (GANs) are also adept at producing realistic images. GANs formulate image generation as a zero-sum game between a generator neural network and discriminator neural network [7]. Two specific architectures are applicable to the style transfer problem: Conditional Adversarial Networks [9], which require pairs of images in two domains as input, and Cycle-Consistent Adversarial Networks [24], which do not require an 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.

3. Methodology

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.

We begin by generating an output image O that is the same dimensions as the content image C with random pixel values. In the first stage, we perform a *global* style transfer based on the technique by Gatys et al. To transfer the style of another reference image S onto C, we update the pixels

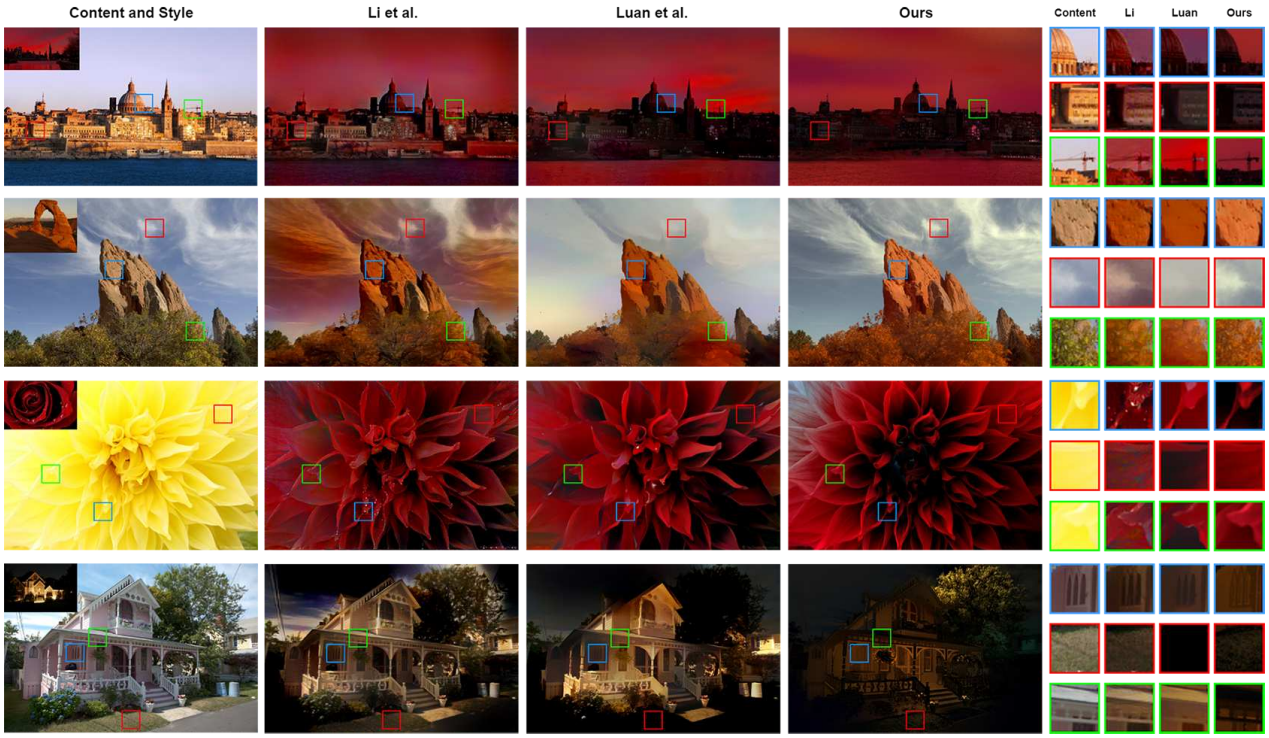

Figure 3: Comparison between our method and the two state-of-the-art photorealistic style transfer algorithms. Details are magnified at the colored boxes for comparison. In particular, our algorithm preserves edges well and smooths out discoloration artifacts. Best viewed digitally.

of O such that it minimizes the following loss function [5]:

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

For a set of 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 flattened 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 feature maps. The scalar weights α and β balance the contribution of the content and style losses to the total global loss.

In the second stage, we perform a *regional* style transfer similar to the technique by Champandard [2] on the output image from the first stage. We now introduce semantic maps of the content and style images as additional inputs, where each map has a pre-determined R distinct regions. For every pixel in the semantic map, we use k-means (setting $k = R$) 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, $M^r_\ell[\cdot] \in \mathbb{R}^{N_\ell \times D_\ell}$, where the elements are 1 if the pixel is in region r or 0 otherwise. We element-wise multiply the masking matrices with the feature matrices, $F^r_\ell[\cdot] = F_\ell[\cdot] \odot M^r_\ell[\cdot]$. Subsequently, we also use the masked feature matrices for Gram matrix calculations. In summary, our regional loss function is:

$$\mathcal{L}_{\text{regional}} = \alpha \sum_{\ell \in L} \mathcal{L}^\ell_{\text{content}} + \beta \sum_{\ell \in L} \mathcal{L}^\ell_{\text{style}}$$
$$\mathcal{L}^\ell_{\text{content}} = \frac{1}{2N_\ell D_\ell} \sum_{r=1}^{R} \sum_{ij} (F^r_\ell[O] - F^r_\ell[C])^2_{ij}$$
$$\mathcal{L}^\ell_{\text{style}} = \frac{1}{2N^2_\ell} \sum_{r=1}^{R} \sum_{ij} (G^r_\ell[O] - G^r_\ell[S])^2_{ij}$$

The use of semantic maps and masking matrices address the content mismatch problem. For each semantic region r, only the pixels in r contribute to the style and content loss due to the masking matrix. This proposition ensures that style is transferred in the corresponding regions.

Our algorithm performs the two stages of optimization sequentially, and repeats for a number of iterations. We choose to not minimize the sum of the global and regional loss functions in one step, but rather alternate the minimization objective in two stages instead. This decision stems from the fact that the style transfer loss functions are non-convex and are highly prone to local minima.

To update the pixels in the generated output image O, we use the Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm (L-BFGS) [14]. Optimization using L-BFGS is computationally expensive due to it being a quasi-

Newton algorithm. The Adam optimizer [12] is a suitable alternative gradient descent algorithm for faster convergence without significant loss of quality in the generated images.

After each update step, we apply a sharpening image filter to the generated image O to clean up remnant distortions and artifacts from the optimization updates. In addition to post-processing, this filtering stage also serves as a constraint on the image search space for the next iteration of the optimization step. In this filtering stage, we can utilize various image processing techniques such as Gaussian blur, high-pass filters, total variation denoising [1], etc. From our experimentation, the most effective post-processing is to apply a guided image filter [8] with the reference content image as a guide. This filter uses the structure and edges in the reference content image to preserve the structure in the generated output image. Since the reference content image is a real-world photograph, this filter transfers the photorealistic structure of a photograph to the generated output image. In our experiments, we apply the filter after the global transfer stage and the regional transfer stage.

3.1. Dataset

Since our approach is a fully optimization-based, there is no need for a training set. For evaluation, we use the evaluation images from Luan et al. [15]. This dataset contains 60 content and style image pairs, along with corresponding semantic segmentation maps for each pair. According to [15], the segmentation maps were automatically generated using DilatedNet [3]. The images are of varying dimension and content, though most are real world photographs.

3.2. Implementation Details

To compute the style transfer loss, we use the VGG-19 network [19] trained on ImageNet [4] as the pre-trained network for feature map extraction. For the style feature layers, we use layers *conv1_1*, *conv2_1*, *conv3_1*, *conv4_1*, and *conv5_1*, setting $\beta_\ell = 1/5$ for each layer. For the content feature layers, we use *conv4_2*, setting $\alpha_\ell = 1$ for this layer. We base our selection of feature layers on the previous work of [13, 15]. For the filtering post-processing stage, we use the guided image filter implementation from MATLAB.

4. Results and Evaluation

Style transfer algorithms are difficult to evaluate, since there are no commonly accepted metrics to measure the faithfulness of the stylistic match between the reference and generated images. Luan et al. [15] conduct a user study to assess the faithfulness of the style transfer and photorealistic quality perceived by human judges. We follow this trend by conducting a user study to qualitatively assess our results. Additionally, we introduce using metrics from the image enhancement and restoration community to numerically assess our results. Since our goal is to produce photorealistic results, we believe metrics that measure image distortion are applicable.

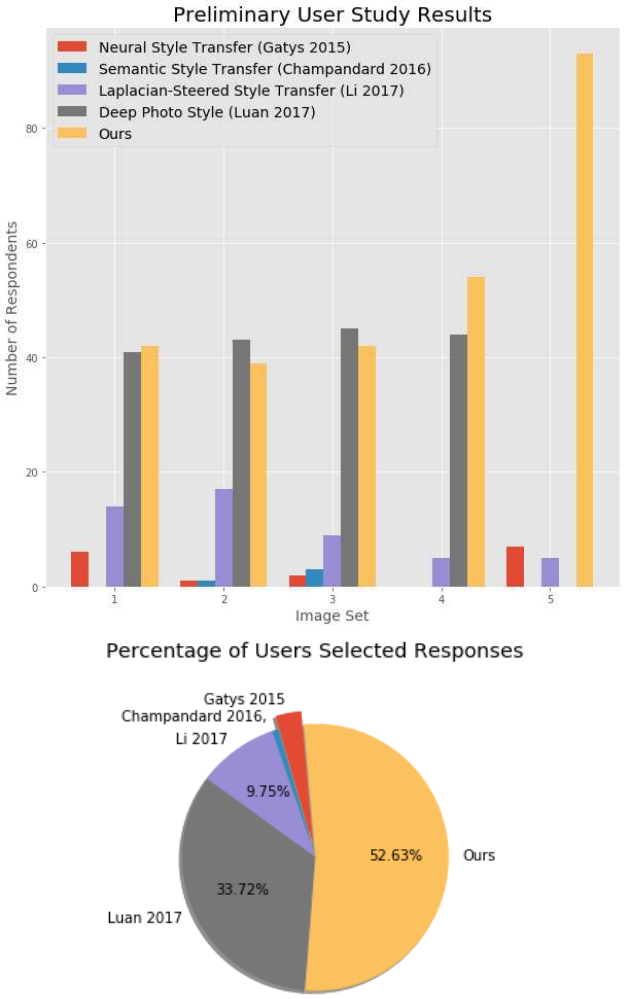

Figure 4: On the top, the number of users that selected images from each technique as the most realistic for each image pair. On the bottom, the percentage breakdown of all user responses.

4.1. Qualitative Evaluation with User Study

To compare our stylized images with previously mentioned techniques, we conduct a user study with 105 human respondents who voluntarily participated online. We randomly select 5 content & style image pairs from the dataset, and generate stylized images using the style transfer techniques from [2] [5] [13] [15] 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

Metric	Content Image	Gatys et al. [5]	Champandard [2]	Li et al. [13]	Luan et al. [15]	Ours
BRISQUE	19.806	29.625	30.433	23.211	21.121	**20.361**
NIQE	2.325	4.236	4.230	3.417	3.169	**2.871**
SSIM	1.0	0.561	0.581	0.579	0.588	**0.656**

Table 1: Mean BRISQUE and NIQE scores of the 60 images generated by all techniques and the reference content image. Mean SSIM is computed between the generated image and the original content image.

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). Across all 5 Image sets, over 52% of users selected our generated images as the most realistic. Our images are comparable those of Luan et al. but our algorithm handles failure cases as well, such as Image set (5). The exact images in Image set (5) are shown in Figure 3 Row 2.

4.2. Quantitative Evaluation with NSS

In perception theory, natural images (i.e. images from the real world) tend to have specific statistical properties [16]. Previously, natural scene statistics (NSS) have been used to evaluate the quality of image de-noising and compression algorithms [20]. Since the goal of this work is to generate photorealistic images, we use various NSS models as an image quality metric.

4.2.1 BRISQUE

In the Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) [16], the algorithm applies pre-determined distortions (e.g. Gaussian noise, blurring) to a dataset of real world images. For each distortion applied to an image, the algorithm keeps a score of the distortion severity. With the distorted images and scores, a support vector regression (SVR) model is trained on the NSS features extracted from said images. During evaluation time, the model predicts a distortion score using an test image's NSS features. BRISQUE returns a non-negative scalar score in the range [0, 100], and lower values correspond to less distortions in the image. We use a pre-trained BRISQUE model from MATLAB to evaluate the images generated by previous techniques our technique for all 60 image pairs in the dataset.

4.2.2 NIQE

In the Natural Image Quality Evaluator (NIQE), the objective is to fit a multivariate Gaussian model on NSS features from a dataset of completely natural images without distortions [17]. During evaluation time, a Gaussian model is fit on the test image's NSS features and the final score is the distance between the Gaussians of the natural image and the test image. NIQE returns a non-negative scalar score without upper bound, and lower values resemble pristine images. We use a pre-trained NIQE model from MATLAB to evaluate the images generated by previous techniques our technique for all 60 image pairs in the dataset.

4.2.3 SSIM

The Structural Similarity Index (SSIM) is a metric for measuring the perceptual quality of a generated image with respect to a reference image [22]. Previous, this metric has been used for assessing the quality of image compression and super-resolution algorithms [21]. The SSIM score of a generated image is in the range of [0,1], where 1 is a perfect structural similarity (e.g. SSIM between the same image).

4.3. Discussion

We use these 3 metrics to quantitatively evaluate our algorithm in different domains. BRISQUE and NIQE estimate the amount of distortions in the generated image and how likely the image is from the real world, respectively. Both of these metrics are important in assessing the photorealistic quality of an image. SSIM measures the perceptual quality of the generated image with respect to the original content image. This is critical since the goal of our algorithm is to produce an image that is a real-world variation of the content image.

In Table 1, we report the mean scores of all 3 metrics for the reference content images and the generated stylized images from techniques we previously referenced. Of all the reported techniques, our algorithm has the best performance for all 3 metrics. Our algorithm has similar performance as the state-of-the-art algorithm by Luan et al. [15], evident in the user study as well. However, we show a strong improvement over the SSIM metric, since our algorithm iteratively preserves structure and edges in the filtering stage. The improvement can be observed in Figure 3 Row 2 and Figure 6 Rows 2 and 3.

4.3.1 Ablation Study

Recall that our algorithm consists of 3 stages: global transfer, regional transfer, and filtering applied after each stage. We perform an ablation study to show the effects of each stage, and the results are shown in Figure 5.

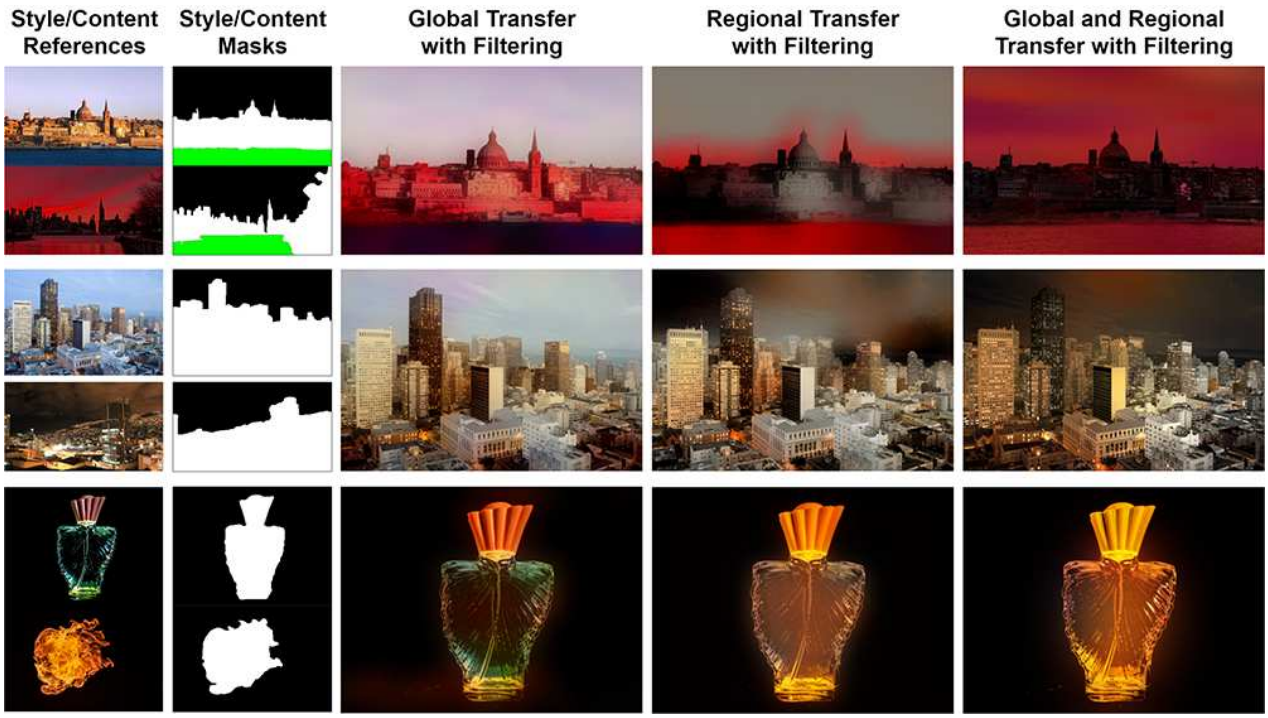

Figure 5: Ablation study performed on each stage of the algorithm. We show that global transfer only results in potential content mismatch, regional transfer only results in artifacts from local minima, and a combination of the two stages are needed to ensure a successful and faithful style transfer.

When applying only global transfer with filtering (shown in the 3rd column), we observe the content mismatch problem we discuss in Section 3. When only using global transfer with filtering, we do not make use of the semantic masks at all. In the first row, we see that the red sky is actually transferred onto the buildings, which is not semantically correct according to the masks. A similar phenomenon is observed for the second row, where the dark sky is not correctly transferred to the blue sky in the image. In the third row, the fire style is transferred to the entirety of the image rather than contained in the bottle.

In the second column, we show the results with only regional transfer and filtering. The primary weakness of this approach is that there are splotches in the generated images. This is due to local minima in the regional loss function.

In the last column, we show the results of our final algorithm that alternates between both the global and regional loss functions. In this combination, the regional loss will constrain the search space for the to potential images without content mismatch, and the global loss will help navigate out of the local minima found in the regional loss function.

All of the aforementioned methods apply the guided image filter after each step. Without image filtering applied, the results should be similar to the previous work in neural style transfer [5] and semantic style transfer [2]. Visually, we can observe that the filtered images contain less distortions than the previous methods where filtering is not used. We draw the conclusion that the guided image filter is the key component in preserving the photorealistic qualities of an generated image.

5. Conclusion

We introduce a novel approach to photorealistic style transfer using a two-stage optimization process and post-processing filtering. Our algorithm maintains both the global and local style from reference input images and segmentation masks. Our approach is fully optimization-based and does not require training a style transfer network on a set of images, nor the use of generative adversarial network architectures.

From a preliminary user study, our technique produces results that are qualitatively comparable to the current state-of-the-art methods but handles additional failure cases. Using natural scene statistics metrics as a quantitative evaluation, our algorithm is shown to create more natural, and realistic images compared to previous methods.

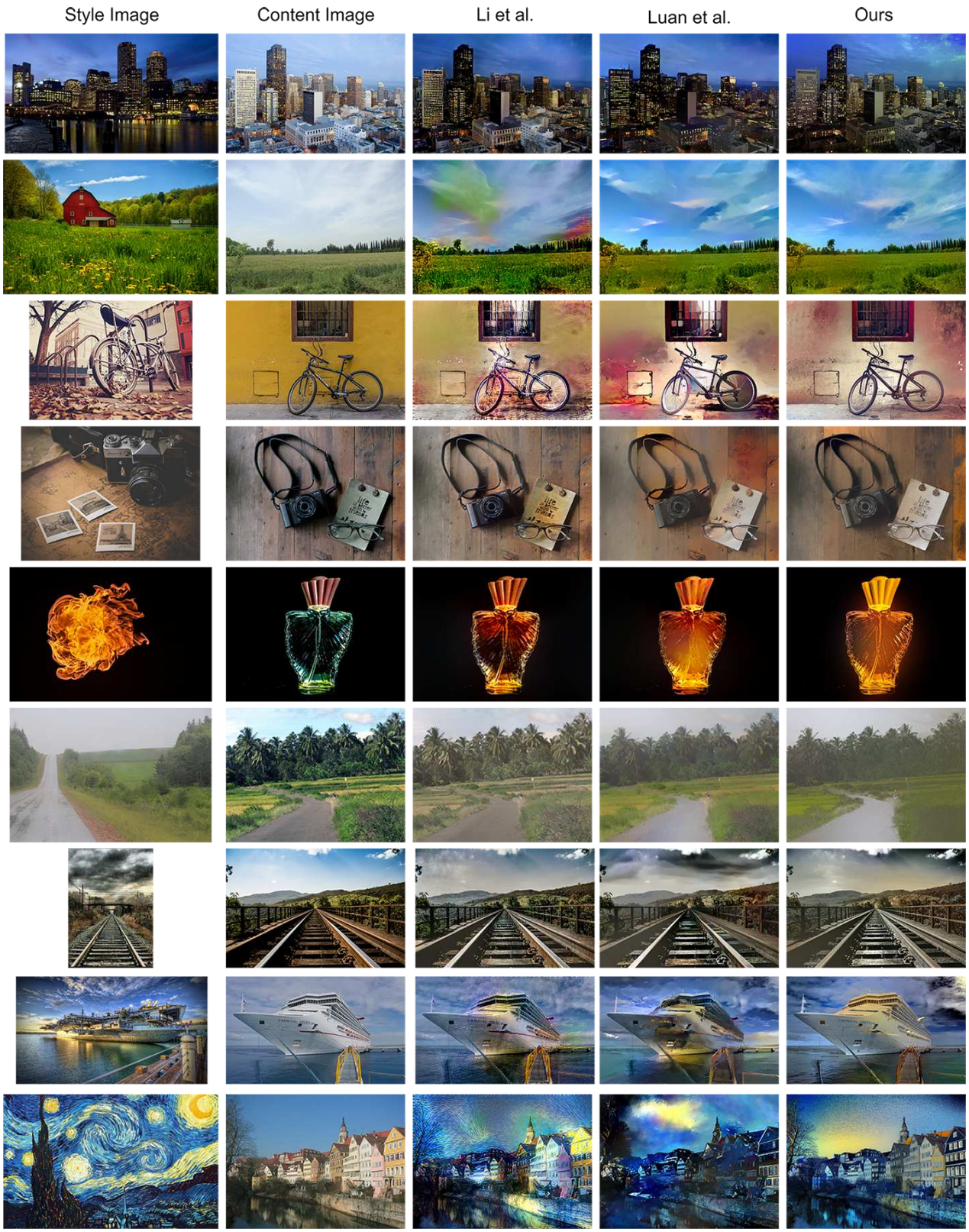

Figure 6: Additional comparisons between our method and the two state-of-the-art photorealistic style transfer algorithms. Best viewed digitally.

References

[1] A. Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20(1-2):89–97, 2004. 4

[2] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016. 2, 3, 4, 5, 6

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 4

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009. 4

[5] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 1, 3, 4, 5, 6

[6] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3985–3993, 2017. 2

[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

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

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 2

[10] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song. Neural style transfer: A review. *arXiv preprint arXiv:1705.04058*, 2017. 2

[11] 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. 2

[12] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[13] S. Li, X. Xu, L. Nie, and T.-S. Chua. Laplacian-steered neural style transfer. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1716–1724. ACM, 2017. 2, 4, 5

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

[15] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. *CoRR, abs/1703.07511*, 2, 2017. 2, 4, 5

[16] A. Mittal, A. K. Moorthy, and A. 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. 5

[17] A. Mittal, R. Soundararajan, and A. C. Bovik. Making a" completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212, 2013. 5

[18] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001. 1

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[20] G. B. Stanley, F. F. Li, and Y. Dan. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *Journal of Neuroscience*, 19(18):8036–8042, 1999. 5

[21] R. Timofte, E. Agustsson, L. Van Gool, M.-H. Yang, and L. Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 114–125, 2017. 5

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[23] H. Zhao, X. Jin, J. Shen, and F. Wei. Real-time photo style transfer. In *CAD/Graphics*, pages 140–145, 2009. 1

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017. 2