

# Depth-Preserving Style Transfer

Xiuming Zhang  
MIT CSAIL

xiuming@mit.edu

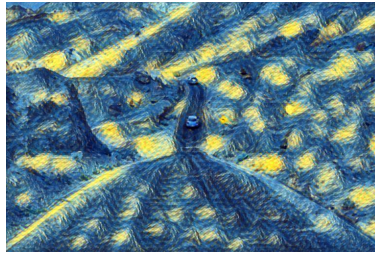
Teammates: Ruizhi Liao & Yu Xia



Style  
*Starry Night*



(A) Scene with large variations in depth



(B) Johnson *et al.*, 2016



(C) Our depth-preserving results

When the input scene exhibits large variations in depth (A), the current state of the art tends to destroy the layering and lose the depth variations, producing a “flat” stylization result (B). This paper aims to address this issue by incorporating depth preservation into the loss function such that variations in depth and layering are preserved in the stylized image (C).

## Abstract

*Style transfer is defined as the process that given a content image and a style image it tries to migrate the style from the style image to the content image. Though it is not clear what the exact definition of style is, pattern transforming and matching are generally accepted.*

*In this work we present a novel method which preserve the depth information of the content image while migrating the style.*

## 1. Introduction

Deep neural networks have gained much popularity in a wide range of computer vision tasks, from low-level tasks, such as image denoising [22] and sharpening, to mid-level tasks, such as keypoint detection [18], to high-level tasks, such as object recognition [12]. Besides supervised tasks with ground truth of training data available (such as scene classification), deep neural networks are also capable of solving more abstract problems, where no ground truth exists for training data. Image style transfer—the process of applying the style of a style input image (*e.g.*, Van Gogh’s *Starry Night*) to a content input image—is one such task, because there exists no “ground-truth stylization” for any

given content input image.

Deep neural networks’ capability of performing image style transfer was first demonstrated by [8], where neural networks are used to both transform images and compute loss. Under this optimization framework, the image transform network iteratively updates its output so that its content loss and style loss are minimized. This approach produces visually pleasant results, but suffers from slow performance due to its iterative nature. To address this issue, Johnson *et al.* [10] trains a residual neural network [9] that, once trained, only needs a feed-forward pass to stylize an input image. Although this method significantly reduces computational burden as compared with [8], it is unable to preserve the content image’s variations in depth during style transfer (teaser figure B), hence destroying the sense of layering.

Although crucial to human perception and aesthetic sense, depth preservation, to the best of our knowledge, has never been accounted for in image style transfer. One of the most common reasons behind this negligence lies in the difficulty of estimating the depth map given a single RGB image. Furthermore, we need the single-image depth estimation module to be fully differentiable in order to incorporate the module into our system for end-to-end training. [1] proposed an hourglass-shape network that meets both re-

quirements: it outputs a depth estimation map from a single RGB image, and it is fully differentiable.

In this paper, we advocate the use of depth loss, defined as the difference in depth perception between the input content image and output stylized image, in the task of image style transfer. We augment the loss network of [10] with a single-image depth estimation network computing how well depth is preserved in the stylized image. Specifically, we define total loss as the sum of style loss, content loss, and depth loss and train the image transform network in the end-to-end fashion. During testing, our image transform network only needs a forward pass to stylize an image, just like [10] does.

Compared with [10], our model produces significantly better results both qualitatively and quantitatively as shown in the experiment section. Qualitatively, we test our model on scenes with large variations in depth. Our model preserves the sense of layer significantly better and produces more homogeneous results than [10]. Quantitatively, we compute the structural similarity (SSIM) index [21] of our stylized images and [10]’s stylized images with respect to the input content image. Indices of our stylized images are unanimously higher than [10]’s, meaning that our model preserves structural similarity better for all test images. Our code is at [github.com/xiumingzhang/depth-preserving-neural-style-transfer](https://github.com/xiumingzhang/depth-preserving-neural-style-transfer).

Our main contributions are twofold. Our first contribution is practical: we demonstrate a simple yet highly effective way of improving the current image style transfer model. Our second contribution points out a possibly promising direction—incorporating loss other than style and content loss (*i.e.*, depth loss in this paper) into the loss function. Applying our idea, one may find that adding, say, shading loss to the network loss function produces more photorealistic results, etc.

## 2. Related Work

The core of our method is incorporating depth preservation losses into the image transformation neural network. Therefore, we review related literature on both neural network-based image style transfer and single-image depth estimation.

### 2.1. Image Style Transfer with Neural Networks

Style transfer can be considered as a more general form of texture transfer, where one transfers texture from one image (style image) to another image (content image). Ideally, semantics of the content image should not be altered in this process. In texture transfer, it is usually the low-level features that are utilized, *e.g.*, in [4].

With the recent prevalence of deep neural networks, researchers started exploring how high-level features extracted by neural networks can be utilized for the task of

style transfer. For instance, Gatys *et al.* perform image style transfer by synthesizing a new image that matches both contents of the content image and styles of the style image [8]. In particular, they extract content representations from the content image and style representations from the style image using the VGG network [17]. Since the VGG network is trained to perform object recognition and localization tasks, the layers deep down the network hierarchy capture object information (*i.e.*, the contents) of the content image and are insensitive to the exact pixel values. Therefore, outputs from these deep layers serve as good content targets that the synthesized image tries to achieve at varying levels of resolution. As for style, they adopt a feature space built on filter responses in any layer of the network [6]. By design, the feature space captures texture information without global arrangement. Finally, they minimize a weighted sum of the content and style loss under a CNN framework, where forward and backward passes are iteratively performed. Building upon this work, the authors recently devised a way of preserve the original colors in the content image [7]. However, the high computational cost still remains as a drawback in [8].

To reduce the computational burden and generate visually similar-quality results, Johnson *et al.* [10] train a feed-forward image transform network to approximate solutions to the optimization problem posed in [8]. In particular, their system consists of a deep residual convolutional neural network (CNN) as the image transform network and the pre-trained VGG network [17] as the fixed loss network. For each style image, the image transform network is trained to apply this style to a content image while minimizing the style and content losses as measured by the loss network. This method produces reasonably good results with low computational cost, but tends to lose the depth variations and destroy layering in the content image as illustrated in the teaser figure. This issue can be addressed by incorporating depth preservation losses into the loss function, as shown later in this paper.

### 2.2. Single-Image Depth Estimation

Deep neural networks trained on ground-truth metric depth data have demonstrated promises in the task of single-image depth estimation [15, 5, 13, 20]. Collecting such ground truth requires specialized cameras, such as Kinect, posing a challenge to large-scale data collections. Although crowdsourcing may seem to be a solution, humans are known bad at estimating absolute depths (which are inherently ambiguous from a single monocular image), but better at judging relative depths [19]. Inspired by this fact, Zoran *et al.* train a neural network to repeatedly judge relative depths of point pairs and interpolate out per-pixel metric depth by solving an optimization problem [23].

Building on [23], a recent work by Chen *et al.* proposes

an end-to-end neural network that takes in a single RGB image in the wild (*i.e.*, taken in unconstrained settings) and outputs pixel-wise depth estimations [1]. Specifically, the deep network follows the “hourglass architecture” recently proposed in [16], which is essentially a series of convolutions and downsampling followed by a series of convolutions and upsampling. Similar to [23], RGB images with relative depth annotations are used as training data. The loss function penalizes large differences in metric depth when the ground-truth relative depth is annotated equal.

### 3. Methods

The overview of our network structure is shown in Figure 1. Compared to Johnson et al. [10]’s work, our structure is featured in having a depth estimation network as part of our loss function. In all, our network is composed of 3 subnets: an image transformation network  $f_W$ , a perceptual loss network  $\phi$  and a depth estimation network  $\delta$ . Similar to Johnson et al., the image transformation  $f_W$  is a convolutional network which produce the output image  $\hat{y}$  given the input image  $x$  by  $\hat{y} = f_W(x)$  (where  $W$  is the weights of the network).

To keep track of the depth information, our loss function is composed of 2 neural networks: the perceptual loss network and the depth estimation network. As mentioned in [10], pretrained convolutional neural networks are able to extract perceptual information and encode semantics which are useful for the loss function. Similarly, a pretrained depth estimation network has already learned to estimate the depth information from the single input image. Therefore we utilize a pre-trained image classification network for the perceptual loss part and a pretrained depth estimation network for the depth loss part. Specifically, our loss function is defined as a weighted linear combination of the content loss  $l_{\text{content}}$ , the style loss  $l_{\text{style}}$  and the depth loss  $l_{\text{depth}}$ .

$$L(\hat{y}, y) = \lambda_1 l_{\text{content}}(\hat{y}, y) + \lambda_2 l_{\text{style}}(\hat{y}, y) + \lambda_3 l_{\text{depth}}(\hat{y}, y)$$

Therefore the training goal is to minimize the expected loss.

$$W^* \leftarrow \arg \min_W \mathbb{E}_{\{x, y\}} [L(f_W(x), y)],$$

where  $\mathbb{E}_{\{x, y\}}$  is the estimation of the expectation via the (training) set  $\{x, y\}$ .

#### 3.1. Depth Loss Function

We make use of depth loss function to measure the amount of depth differences between the input image  $x$  and the output image  $\hat{y}$ . Ideally, the output image should have similar depth features with that of the input. Rather than

capture the per-pixel differences of the feed-forward outputs we capture the high level features from the depth estimation network. More specifically, we define the depth loss function  $l_{\text{depth}}$  as the 2-norm of the feature vectors (from selected layers)

$$l_{\text{depth}}(\hat{y}, y) = \sum_{i \in I_\delta} \frac{1}{N_i(\delta)} \|\delta_i(\hat{y}) - \delta_i(y)\|_2^2$$

, where  $N_i(\delta)$  is the normalizing factor for the  $i$ -th layer in  $\delta$  and  $\delta_i(y)$  indicates the feature vector on the  $i$ -th layer if  $y$  is feeded as the input to the network  $\delta$ . The layer set  $I_\delta$  means the set of (high-level) layers we want to extract features from. The motivation for a high level depth loss function is that we want to encourage the output from  $f_W$  to be similar to the content image from the depth perspective but we don’t want their depth estimation to be exactly the same. There are several reasons for such a motivation: firstly, the estimations of depth from the network  $\phi$  are not necessarily accurate which make it meaningless to pursue a (per-pixel) exact match on the depth estimation. Secondly, we need to allow the image transformation network  $f_W$  to perceptually transform the image which might involve changes of shapes, places and lines. Again it is not promising to propose a per-pixel loss, which reduce the chances of such transformations. Thirdly, as argued in [10], perceptual losses are more robust and stable than the per-pixel losses.

#### 3.2. Content Feature Loss Function and Style Loss Function

For content feature loss function  $l_{\text{feat}}$  and the style loss function  $l_{\text{style}}$ , we briefly recall the explanation from [10]. As one of the main contributions in Johnson et al.’s paper,  $l_{\text{feat}}$  and  $l_{\text{style}}$  are both measures of differences of high-level features.  $l_{\text{feat}}$  captures the distances in respect of perceptual features between the content target  $y_c$  (*i.e.*, the input image  $x$ ) and the output image  $\hat{y}$ . Similarly, as proposed from Gatys et al. [8],  $l_{\text{style}}$  captures the distances between the style image  $y_s$  and the output image  $\hat{y}$ . Therefore

$$l_{\text{feat}}(\hat{y}, y) = \sum_{i \in I_\phi} \frac{1}{N_i(\phi)} \|\phi_i(\hat{y}) - \phi_i(y)\|_2^2,$$

and for style loss we use the Frobenius norm of differences of the Gram matrices of  $\hat{y}$  and  $y_s$ .

$$l_{\text{style}}(\hat{y}, y_s) = \sum_{i \in I_\phi} \frac{1}{N_i(\phi)} \|G_i^\phi(\hat{y}) - G_i^\phi(y_s)\|_F^2.$$

### 4. Experiments

#### 4.1. Training Details

Microsoft COCO dataset [14] (containing around 80K images) was used for training our depth-preserving style transfer networks. Each training image was resized to  $256 \times 256$ .

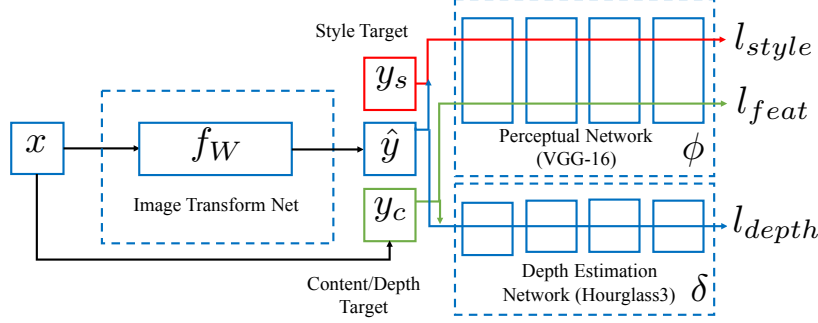


Figure 1: Network Structure Overview

Maximum iterations were set to be 40000, and a batch size of 3 was applied. These settings gave roughly 1.5 epochs over all the training data. The optimization was based on Adam [11] with a learning rate of  $1 \times 10^{-3}$ . No weight decay or dropout was used. The training was implemented using Torch [3] and cuDNN [2]. Each style training took around 7 hours on a single GTX Titan X GPU. As for the target losses, the content target loss is computed at VGG network layer *relu2\_2*, the style target loss is computed at VGG network layers *relu1\_2*, *relu2\_2*, *relu3\_3* and *relu4\_3*, and the depth target loss is computed at the output layer of Chen *et al.* network [1].

## 4.2. Qualitative Results

In this section we present some results regarding the quality of our outputs. Generally, Structural Similarity Index is a method to measure the similarity between 2 images. It can be seen as a quality measure if the other image is considered truth (or perfect). For our case, we need to calculate the index between the original input and style transferred images. We measured the Structural Similarity Index against the original input of our results and the results from Johnson *et al.* [10] on all the input images we tested with the style files provided by [10]. As shown in Figure 2, our results has a generally higher Structural Similarity Index than those from [10], which indicates that our method preserved much more structural information.

## 5. Discussion and Conclusions

Some might argue that if the weight for the style target loss in the Johnson *et al.* [10] network is decreased, we could get similar results. But this is not the case. Note that the content target loss term is computed as the distance of the feature representations in the VGG network [17], which was designed and trained for object recognition. The backgrounds in the content images, for instance, skies, roads, and lawns, are hard to be fully represented in those features. Also, the depth of an “object” (for example, the road in the teaser image) may change a lot in an image, and some

pixel-level (or superpixel-level) loss should be included to preserve stereopsis of the transferred images. Therefore, an extra depth loss is necessitated in our problem setting.

The intuition of why this depth loss net works is that style is put onto the content images layer by layer, not equally, when including the depth loss. As we can see from the outputs of Johnson *et al.* [10] network, the style is transferred almost equally to different regions of the content images. This may destruct stereopsis of the transferred images. The depth loss network we employed here, however, penalizes “equal style transfer” across the content images.

Some textures in our transferred images are also preserved as a by-product, for example, the rock texture and the grass texture in Figure ?. Again, this is due to that under the depth loss penalization, background features can be captured and preserved. This helps preserve textures of background objects.

In conclusion, the state-of-the-art work in style transfer [10] utilized the VGG network to capture perceptual features from content images, and they showed significant improvement of transfer speed compared to the previous work using pixel-level losses. However, the perceptual features extracted from layers in the VGG network can not fully represent background information. Especially, the depth of content images could not be preserved in some of their outputs. In this work, by combining the depth loss network with the perceptual losses, we demonstrated that our style transfer can preserve depth and stereopsis of the original content images, without losing the benefit of transferring speed.

## References

- [1] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. *arXiv preprint arXiv:1604.03901*, 2016.
- [2] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

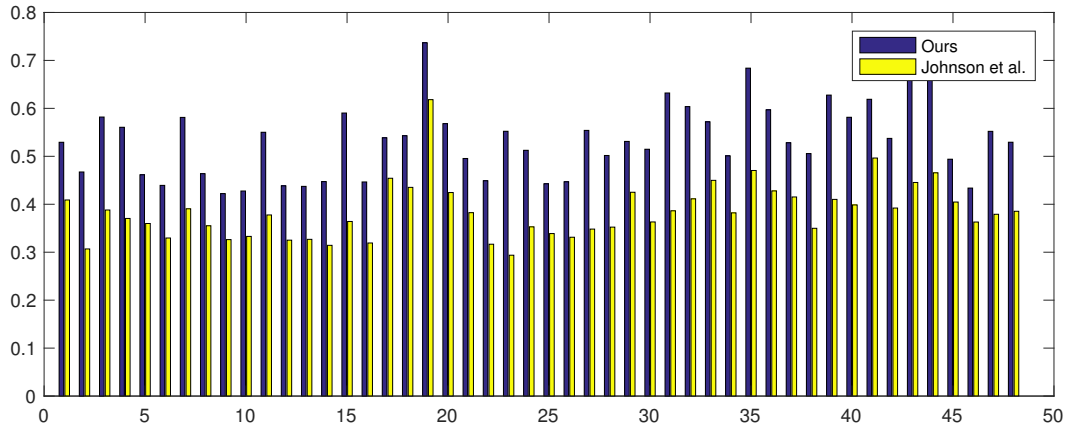


Figure 2: Comparison of Structural Similarity Indices Against the Input Image Between Ours and Johnson et al. [10]

- [3] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.
- [4] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001.
- [5] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.
- [6] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 262–270, 2015.
- [7] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.
- [8] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.
- [10] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *arXiv preprint arXiv:1603.08155*, 2016.
- [11] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [13] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1119–1127, 2015.
- [14] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
- [15] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.
- [16] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*, 2016.
- [17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [18] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.
- [19] J. T. Todd and J. F. Norman. The visual perception of 3-d shape from multiple cues: Are observers capable of perceiving metric structure? *Perception & Psychophysics*, 65(1):31–47, 2003.
- [20] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. Yuille. Towards unified depth and semantic prediction from a single image. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2800–2809. IEEE, 2015.
- [21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
- [22] J. Xie, L. Xu, and E. Chen. Image denoising and inpainting with deep neural networks. In *Advances in Neural Information Processing Systems*, pages 341–349, 2012.
- [23] D. Zoran, P. Isola, D. Krishnan, and W. T. Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 388–396, 2015.