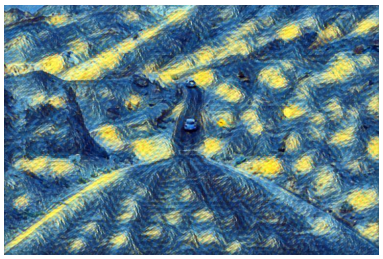
Depth-Preserving Style Transfer

Xiuming Zhang
Massachusetts Institute of Technology
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

convolutional neural network (CNN)

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 [2].

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 [6]. In particular, they extract content representations from the content image and style representations from the style image using the VGG network [11]. 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 [4]. 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. Build-

ing upon this work, the authors recently devised a way of preserve the original colors in the content image [5]. However, the high computational cost still remains as a drawback in [6].

To reduce the computational burden and generate visually similar-quality results, Johnson *et al*. [7] train a feed-forward image transform network to approximate solutions to the optimization problem posed in [6]. In particular, their system consists of a deep residual CNN as the image transform network and the pretrained VGG network [11] 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 [9, 3, 8, 13]. Collecting such ground truth requires specialized cameras, such as Kinect, posing a challenge to large-scale data collections. Although crowd-sourcing 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 at judging relative depths [12]. 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 [14].

Building on [14], 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 [10], which is essentially a series of convolutions and downsampling followed by a series of convolutions and upsampling. Similar to [14], 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

4. Experiments

5. Discussion & Conclusion

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] 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.

[3] 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.

[4] 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.

[5] L. A. Gatys, M. Bethge, A. Hertzmann, and E. Shechtman. Preserving color in neural artistic style transfer. *arXiv preprint arXiv:1606.05897*, 2016.

[6] 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.

[7] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *arXiv preprint arXiv:1603.08155*, 2016.

[8] 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.

[9] 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.

[10] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*, 2016.

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

[12] 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.

[13] 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.

[14] 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.