Style Transfer Incorporating Depth Perception and Semantic Segmentation

XU, YINGHAO YANG, YUXIN ZHANG, YUN

CS 231N PROJECT, STANFORD UNIVERSITY

Introduction

- Current vanilla neural style transfer trained on deep convolutional neural networks discard most depth information, which can be an critical to aesthetics.
- **Gaty's model:** optimization method
- **Johnson's model:** train a feed-forward network that minimizes both style and content perceptual differences.
- **Our Task:** We explore ways to incorporate depth information into existing style transfer algorithm based on Johnson's work to build a fast, clear, and operationally inexpensive style transfer model.

Problem Statement

- **Objective**: In this project, we investigate ways to extend the perceptual loss functions in Johnson's paper and train a model that realizes image style transfer with depth information better preserved, with only a single forward pass of input image.
- **General Approach:** Train an image transformation network to transform input images into output images which minimizes three perceptual losses (style loss, content loss and depth loss). The loss networks remain fixed during training process.
- **Evaluation:** Depth preservation and aesthetic beauty.

Datasets & Pretrained Models

Trained Model: Image Transformation Net

- Style transfer networks train on the MS-COCO dataset

The 80k training images in the dataset we use has been resized to 256 x 256 and we train with a batch size of 4 for 20k iterations, giving roughly 1 epoch over the training data.

Figure 1: Microsoft COCO sample training datasets

Obtained Models:

- **Depth Prediction:** "Res-Net-UpProj" trained on NYU Depth v2 & Make 3D

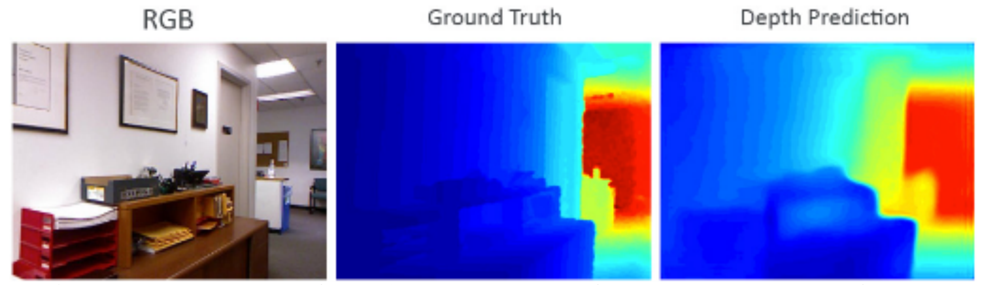

Figure 2: Example of NYU Depth2 v2 Net Results

Datasets & Pretrained Models continued

Obtained Models:

- **Loss network :** VGG-19

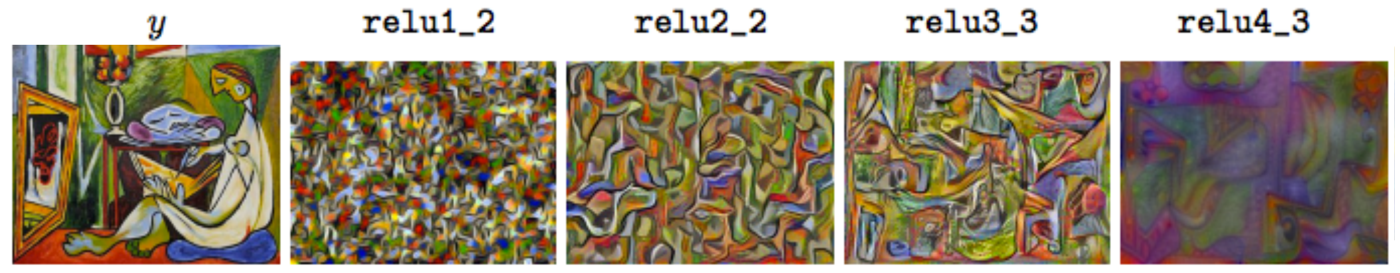

Figure 3: Output of different layers of VGG-16

Models and Algorithms

- **Style Transfer trained with Style, Content and Depth Loss**

We trained our fast neural depth perceptive style transfer model using following architecture:

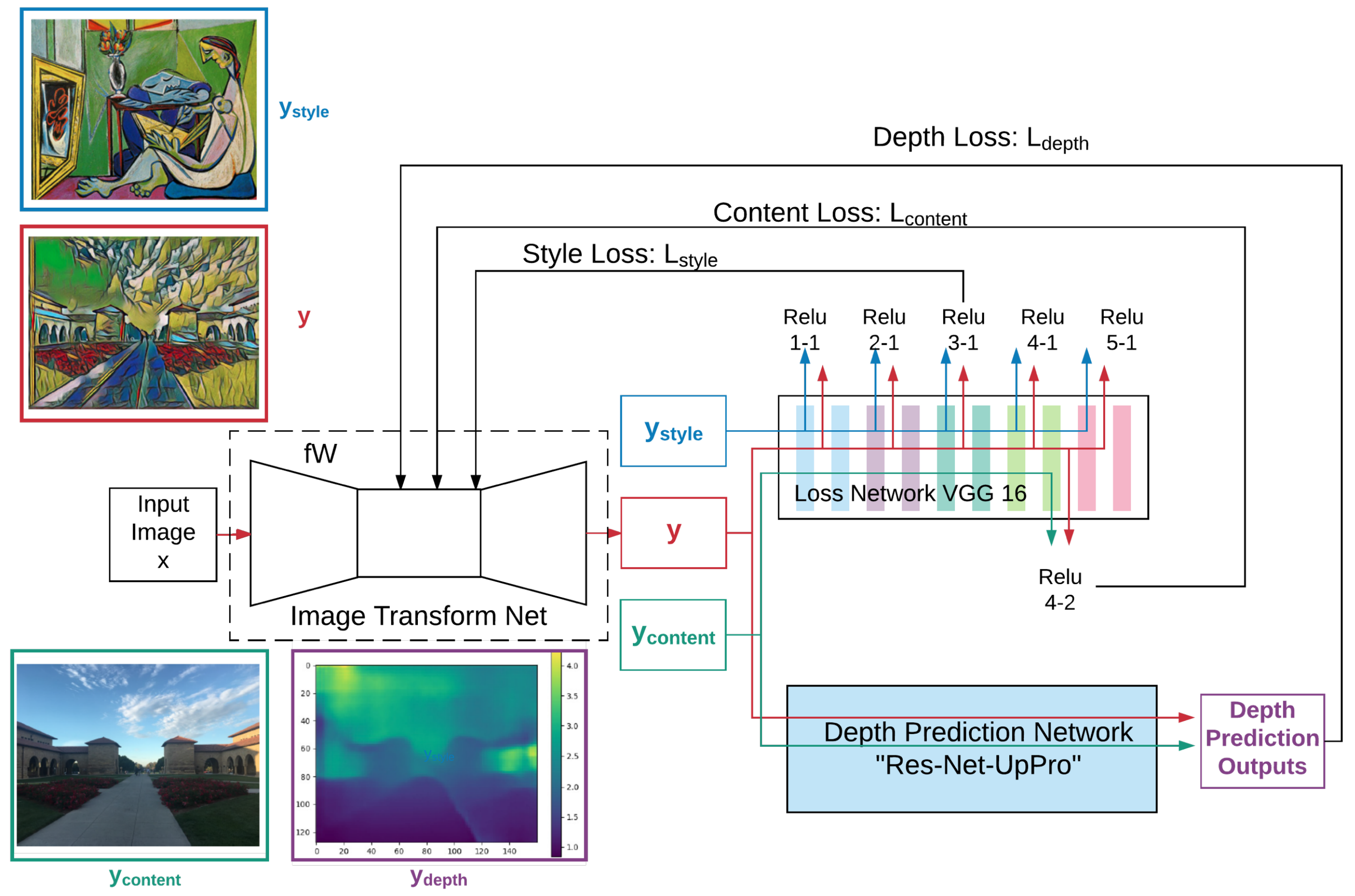

Figure 4: System Architecture Overview

Output With Depth

- **Perceptual Loss Functions**

We redefined the reconstruction loss functions between y and $\hat{y}$ as follows:

$$l^{\Phi,j}_{style}(\hat{y}, y_{style}) = \|G^{\Phi}_j(f_W(x)) - G^{\Phi}_j(y_{style})\|^2_F$$

$$l^{\Phi,j}_{content}(\hat{y}, y_{content}) = \frac{1}{C_i H_i W_i}\|\Phi_j(f_W(x)) - \Phi_j(y_{content})\|^2_2$$

$$l^{\Psi,j}_{depth}(\hat{y}, y_{content}) = \frac{1}{C_j H_j W_j}\|\Psi_j(f_W(x)) - \Psi_j(y_{content})\|^2_2$$

Experiment and Results

- **Experiment:**

In our experiment, we trained the image transformation net for 1 epoch, which is roughly 20k iterations, with batch size 4, learning rate 1e-3, style loss weight 1e2, content loss weight 7.5e0 and depth loss weight 1e2. We use Adam optimizer to find an output image that minimizes the perceptual loss.

Experiment and Results continued

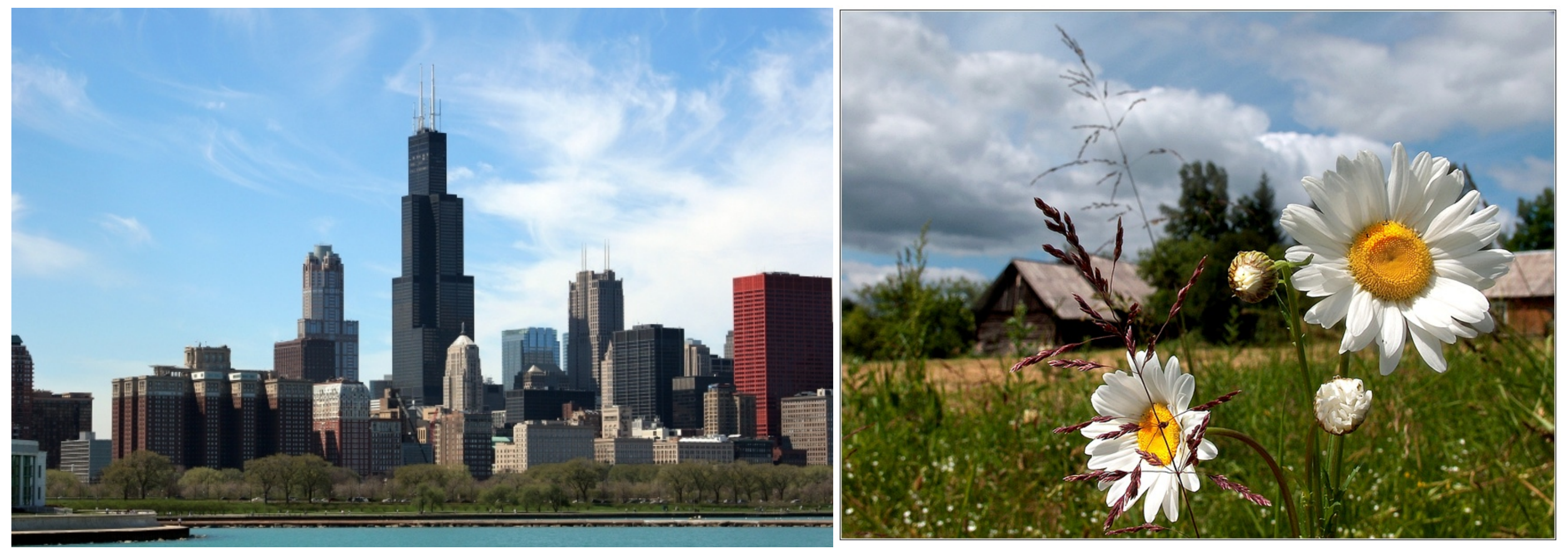

Figure 5: Input images

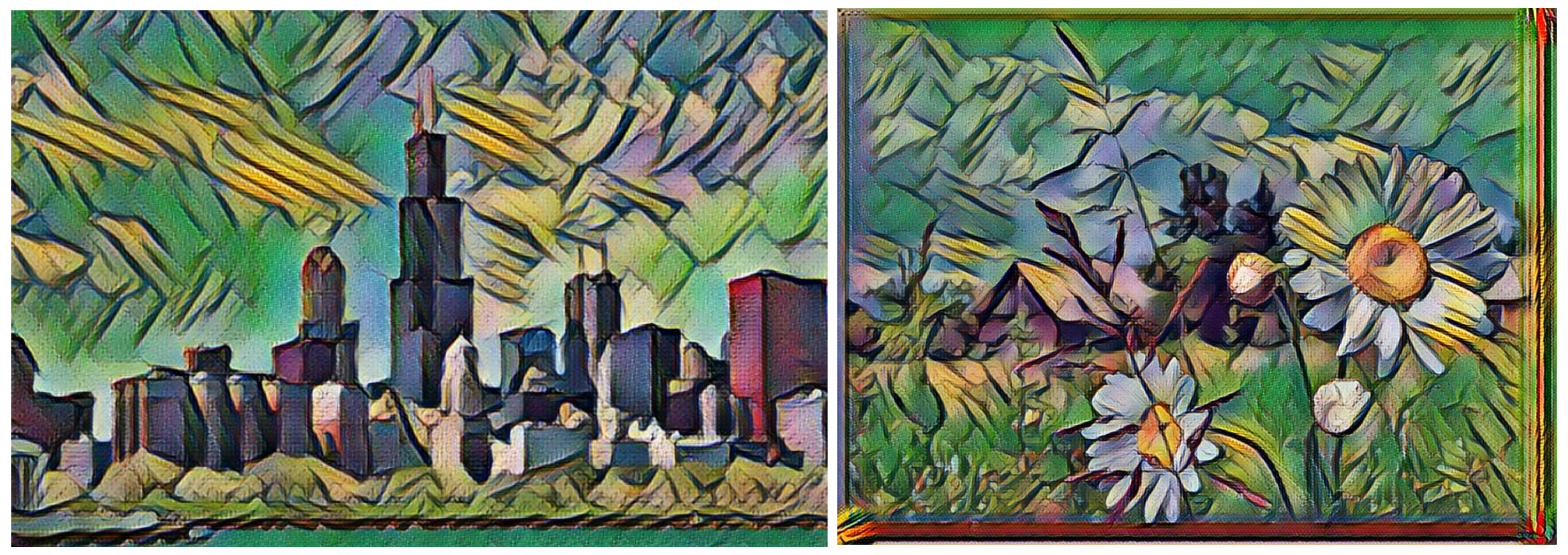

Figure 6: Original output images

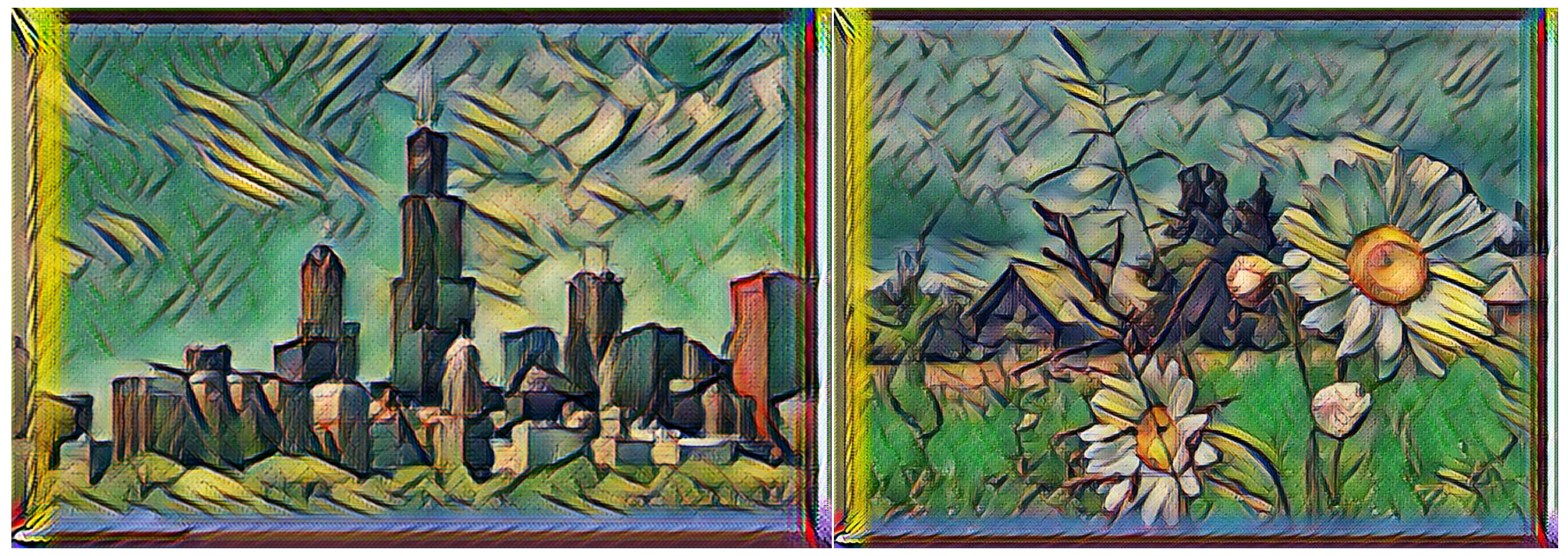

Figure 7: Output images with depth preservation

Analysis and Future Work

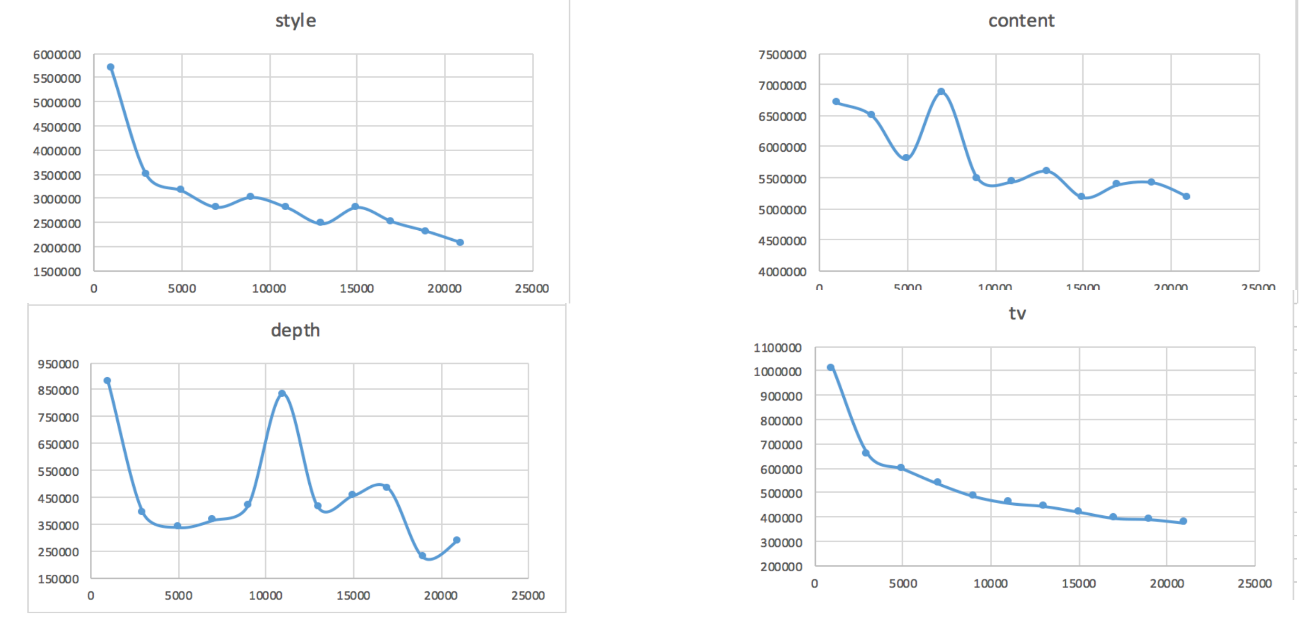

Figure 8: Different losses vs. iteration number

- Will further fine tune our model to obtain better outcomes.
- Will incorporate segmentation information into our network.