Style Transfer with Prominent Stereoscopic Impression Based on Residual Attention Capsule Network

ChenYu, YanBohan, WangQianzhen

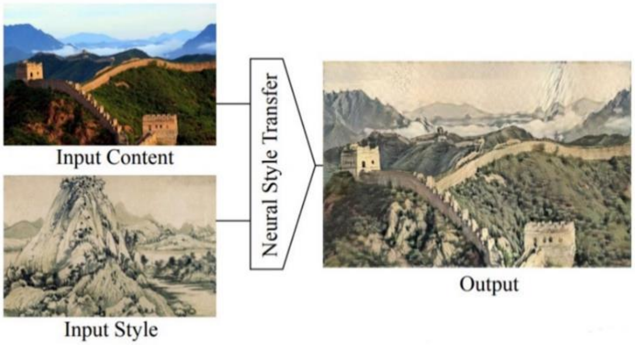
Abstract

As time goes on, the time efficiency of neural network style transfer algorithm is constantly improved and the limits are gradually being broken, which is exciting. **However, the quality of output images did not get enough attention because of its inherent measurement difficulties.**

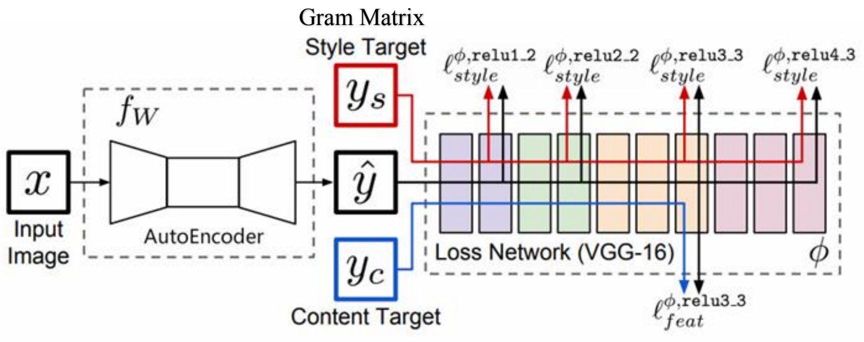
In traditional style transfer tasks, artistic images of two-dimensional composition are often used as style labels, which can result in the planarization of images of three-dimensional composition. In this project, we modify the representation of the style matrix to improve the quality of output and introduce attention mechanisms as well as capsule layers to create a prominent stereoscopic impression.

I: Definition

Neural style transfer refers to combining the content of one image with the style of another image and then generate a new image with both characteristics.



II: Related Work


Gram Matrix
Style Target
y_s
$\ell_{style}^{\phi,relu1\_2}$
$\ell_{style}^{\phi,relu2\_2}$
$\ell_{style}^{\phi,relu3\_3}$
$\ell_{style}^{\phi,relu4\_3}$
f_W
x
Input Image
AutoEncoder
$\hat{y}$
y_c
Content Target
Loss Network (VGG-16)
ϕ
$\ell_{feat}^{\phi,relu3\_3}$

Generative neural networks are widely used in semantic segmentation, protein structure prediction and other generative tasks, because of their strong creative ability. We use a special generative neural network—autoencoder to achieve the image style transformation task. A pretrained neural network for image classification can be used to define perceptual loss functions that measure perceptual differences in content and style between images. By minimizing the perceptual loss between output and input images, the autoencoder can generate an image with artistic style and input image content.

III: Infrastructure & Approach

1. The usual style representation method uses gram matrix on channel dimension. Each element of the matrix is the inner product between two features, which may loss the spatial style of the feature maps. Therefore, we use gram matrix again instead of taking the inner product.

2.we introduce capsule layers into the deeper level, capsule layer can keep more information and the dynamic routing mechanism can help the network to integrate hierarchical relationship. On the other hand, our model imitates the human brain's attention mechanism to make the features of each object more distinct.

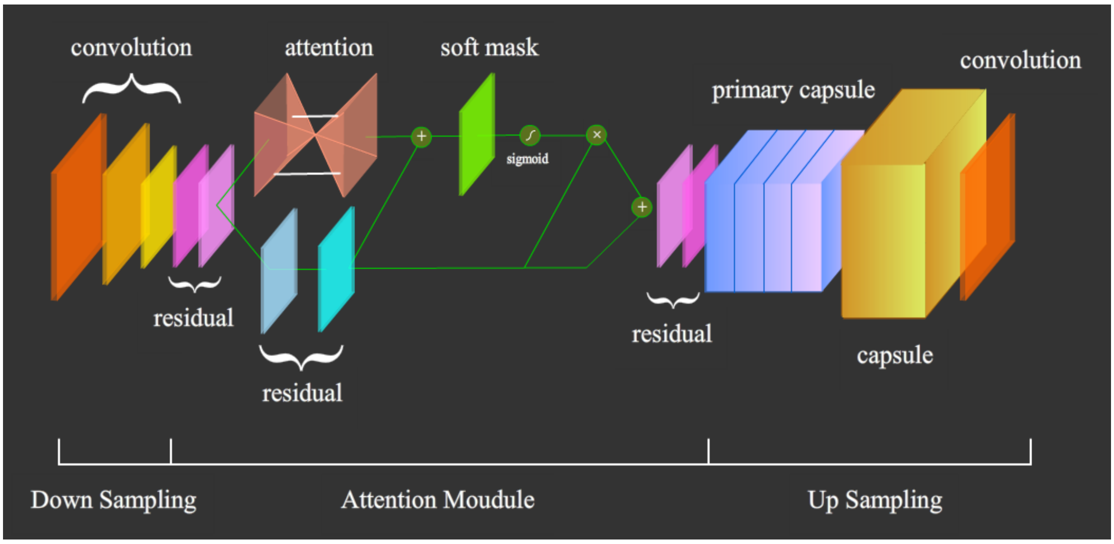
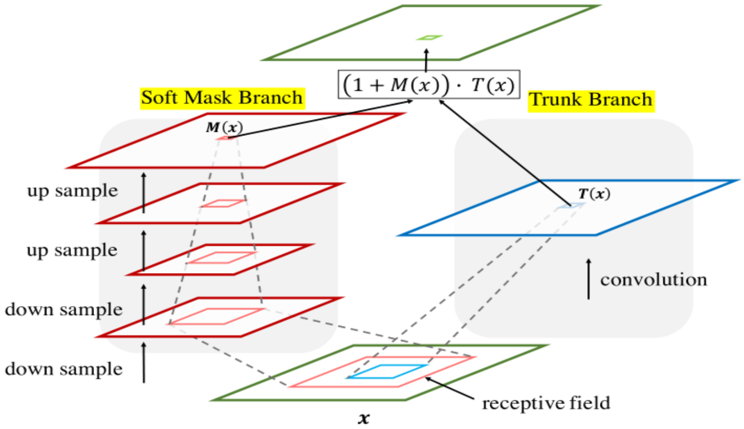


Fig.1 The autoencoder is used to transform images' style, in our final version, there are three parts in it: down sampling, attention module and up sampling.

1. Down sampling

In this part, we use three convolution blocks to extract the primary features of the input image. Each convolution block contains four parts: 1. ReflectionPad2d: it is helpful to reduce the marginalization effect in the image reconstruction task; 2. Conv2d, 2-dimensional convolution operation; 3. InstanceNorm2d, we use instance normalization instead of batch normalization, which is more appropriate for style migration tasks; 4. ReLU, nonlinear activation function. Through this part, the features' channels are changed from 3 to 128.

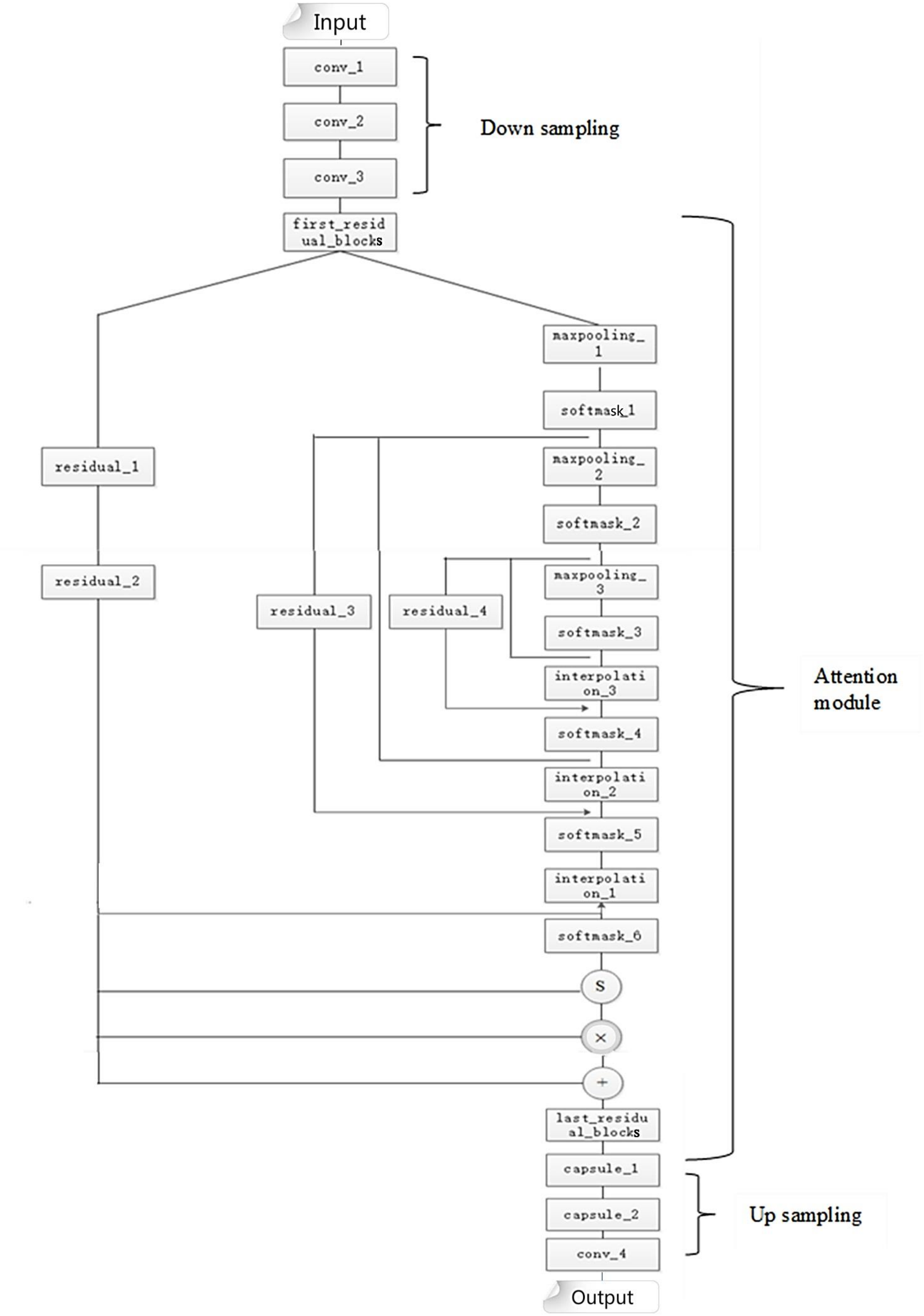
2. Attention Module



At the middle layer of the autoencoder, we use the attention module to extract salient features of objects. The attention module has two parts, which are trunk branch and soft mask branch. For the trunk branch, we use two residual blocks. For the soft mask branch, it has the bottom-down and top-up structure, which imitates the brain's attention mechanism. We use the max pooling to execute down-sampling, and use linear interpolation to execute up-sampling. For more specific processes, please refer to the detailed network structure diagram provided as Fig.2. Through attention module, the number of features' channels are not changed.

3. Up-sampling

In the last part, there are two deconvolution capsule layers and one convolution layer in total. At the beginning, the primary capsule layer groups the original 128 scalar features into two pairs and takes 64 vector features as input, and outputs 4 vector features of length 16. On the other hand, the next capsule layer converts the feature into a vector feature of length 32, which contains pixel information. Finally, a convolution layer samples this huge vector feature, extracting the 3-channel scalar.



IV: Literature Review

Style transfer methods can be divided into two categories: “Descriptive Neural Methods Based on Image Iteration” and “Generative Neural Methods Based on Model Iteration”. The first method achieves image style migration by directly updating image pixels iteratively. The second method firstly optimizes the generation model iteratively, and then generates stylized images.

Some important papers:

1.Descriptive Neural Methods Based on Image Iteration

“Image Style Transfer Using Convolutional Neural Networks” 2016 CVPR

2.Generative Neural Methods Based on Model Iteration

- **One style, one model**

“perceptual losses for real-time style transfer and super-resolution” 2016 ECCV

- **Multiple styles for one model**

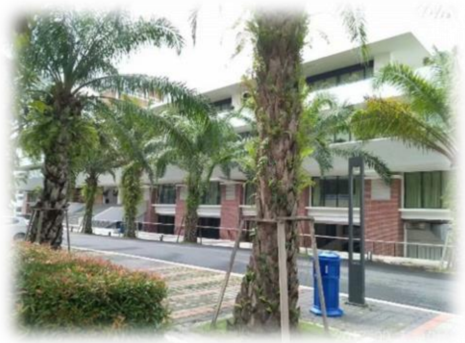
“A Learned Representation for Artistic Style” 2017 ICLR

- **Any style of a model**

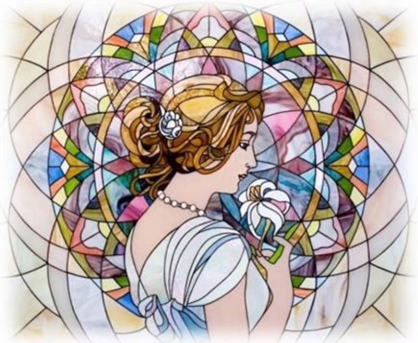
“Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization” 2017 ICCV

V: Error Analysis

Our work based on *“perceptual losses for real-time style transfer and super-resolution”*. In fact, we did many different experiments with the same loss weights and iteration number to find the best way.

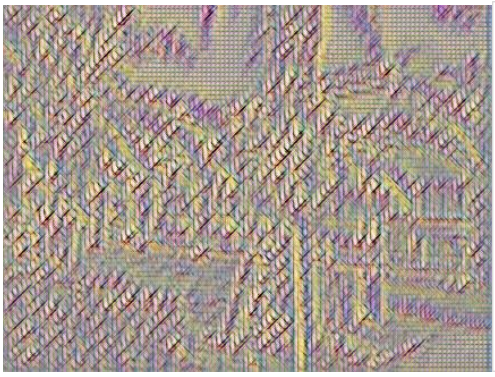


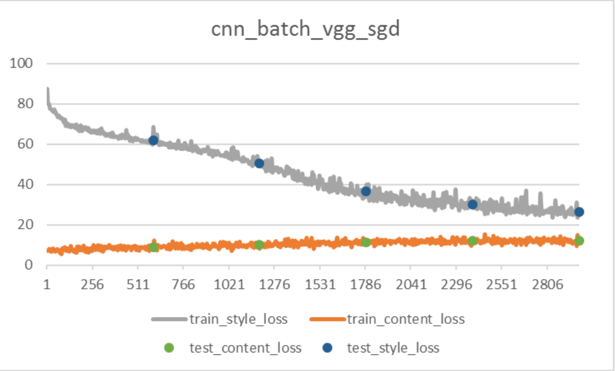
Content



Style

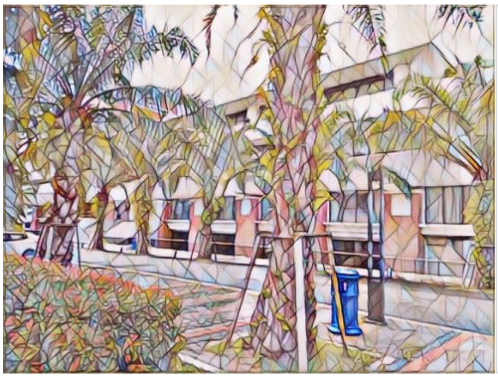
Original network + Batch normalization + SGD

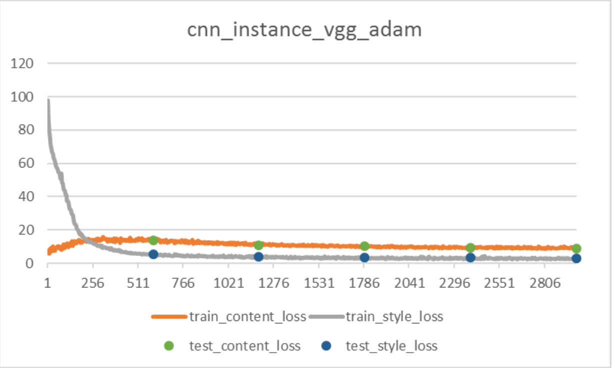




In this experiment, mainly based on the original paper, we found that the network convergence is slow and the final effect is not good. We think SGD optimization speed is very slow, Adam is more suitable. In addition, compared with instance normalization, batch normalization is not suitable for our tasks, which may lead to more efforts needed to achieve good training results.

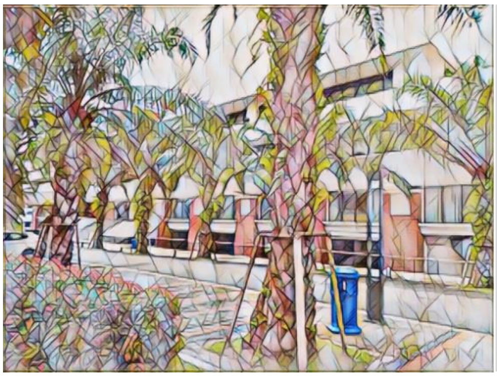
Original network + Instance normalization + Adam

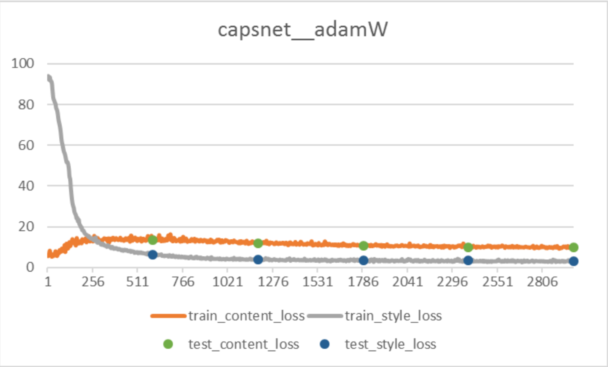




In this instance, the results of the output images were promoted, which proved that the changes were correct.

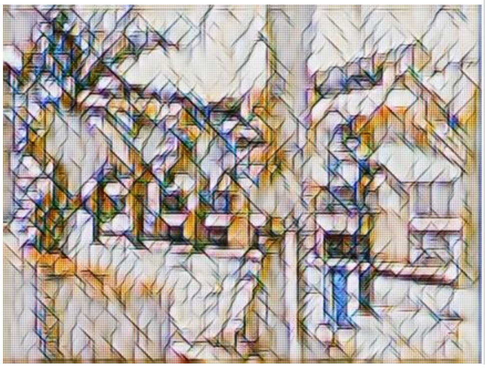
Attention Capsule Network + Instance normalization + Double Gram + AdamW

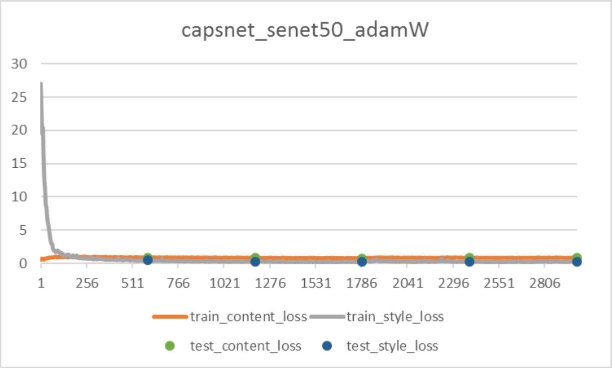




The result of this experiment is the most satisfactory. The whole composition is closer to the original picture. The features of the objects are well highlighted, which leads to a more shocking visual effect.

VGG → SeNet50



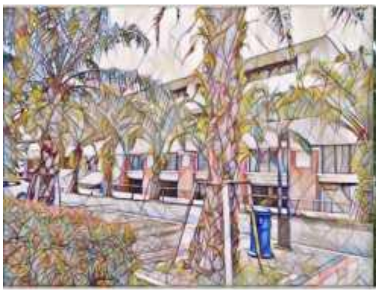


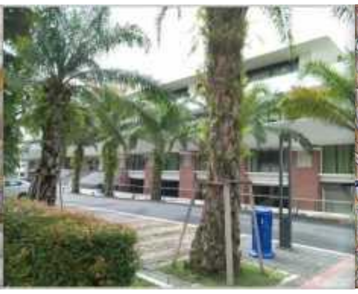
In addition to previous three experiments, we also tried to replace the loss network to Senet50. However, the effect of that experiment was not good. We conject that it is because the location of the extraction layer and the loss weights were inappropriate. Due to limited time and computing resources, we did not make more attempts to adjust these hyper-parameters.

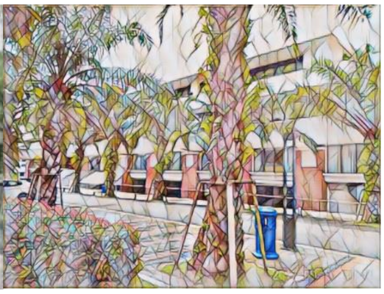
The differences of results can be observed from two perspectives.

Perspective I

Original network output Input Our network output



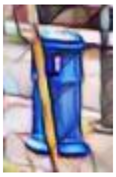


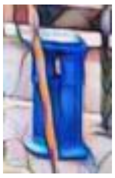


Perspective II

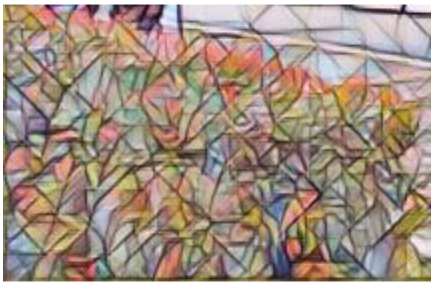
1. Planarization

Original output Input Our output



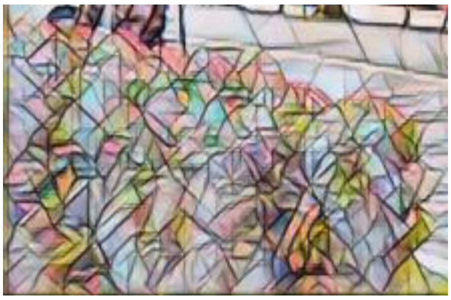


2. Color differences.



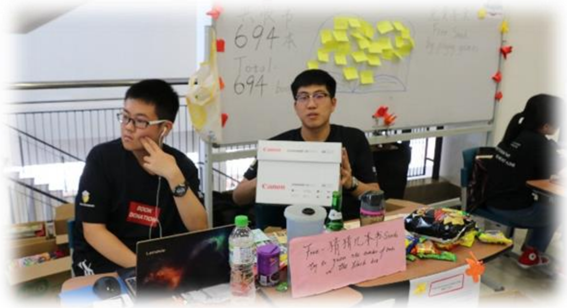
Original network output

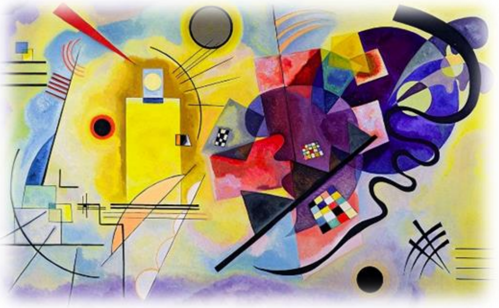
Input image



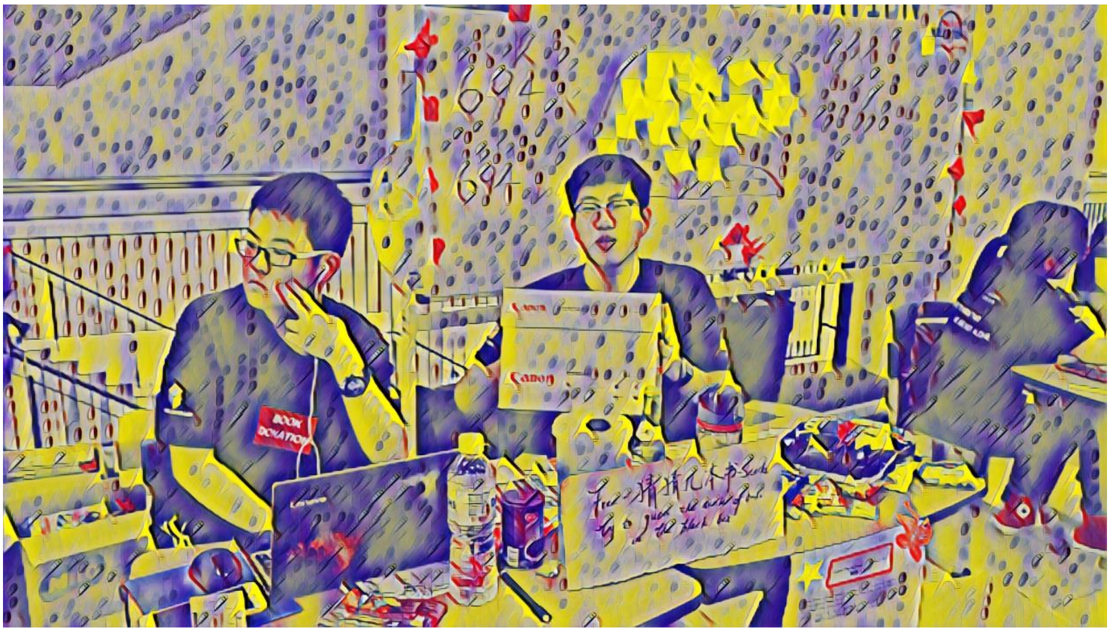
Our network output

Another comparison

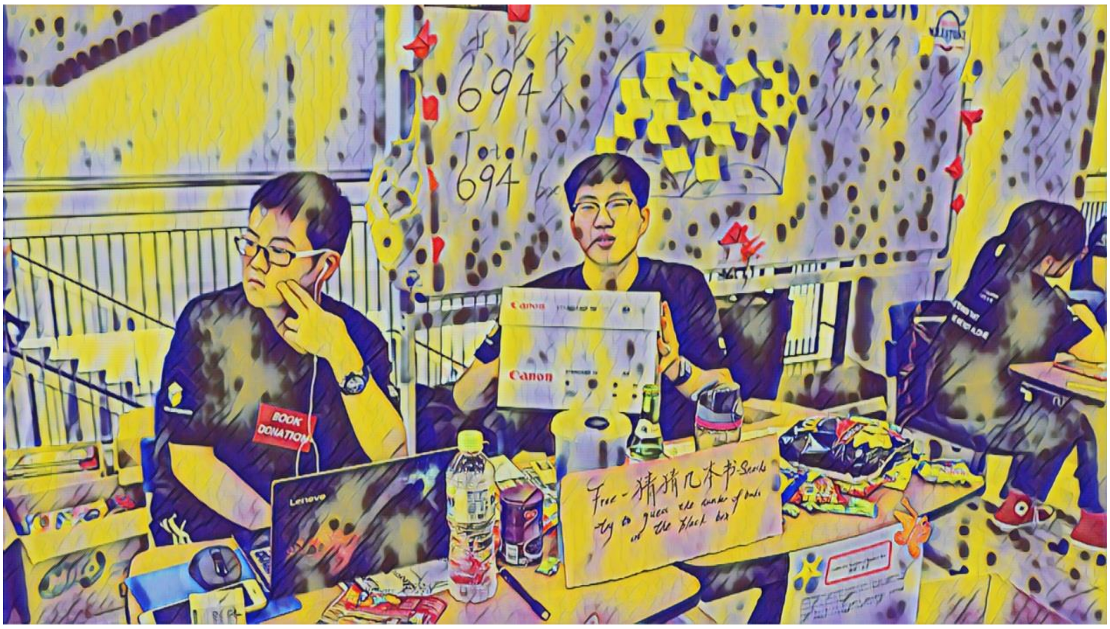




Original network output



Our network output



VI: Conclusion

In this project, we propose a new transformation network model, which can generate high quality output images thanks to the attentional soft mask and capsule layers. Experimental Results.

VII: Reference

1. “A Neural Algorithm of Artistic Style” https://arxiv.org/pdf/1508.06576.pdf
2. “Dynamic Routing Between Capsules” https://arxiv.org/abs/1710.09829
3. "Fixing Weight Decay Regularization in Adam" https://arxiv.org/abs/1711.05101
4. “Instance Normalization: The Missing Ingredient for Fast Stylization” https://arxiv.org/abs/1607.08022
5. “Perceptual Losses for Real-Time Style Transfer and Super-Resolution” https://arxiv.org/abs/1603.08155
6. “Residual Attention Network for Image Classification” https://arxiv.org/pdf/1704.06904.pdf
7. “Squeeze-and-Excitation Networks” https://arxiv.org/abs/1709.01507