CS216 Project
Analysis of Artistic Image Synthesis and Image Mosaics

Yifan Tian 78921267

1 Project Introduction

In this project, I finish artistic style transfer. Based on this, I develop a new way to do image mosaic. Below are outlines of my studies.

1. **Style synthesis and artistic image synthesis**
   In this part, I mainly repeated part of work in other articles. Besides repeating, I also do some study of the parameters and do some new explorations like constructing images using two styles, combining two contents.
   - Artistic image synthesis applied to a house image with famous artwork the "The Starry Night", and "The Scream".
   - Analysis of effect in terms of sweeps and parameters. I do some experiments by using different numbers of steps in optimization and different parameters in loss functions.
   - Artistic image synthesis with two styles.
2. **Image reconstruction with two contents**
   Try to use two contents in content reconstruction to see if it can give an interesting result.
   - Do two-content reconstruction using contents from house image and pebbles image. These two are images that have no similarities.
   - Do two-content reconstruction using contents of faces from two famous figures. These two faces are similar in terms of their spatial arrangement.
3. **Multi-resolution Mosaic**
   This part is mainly my new research and interesting part. This is based on my understanding of content reconstruction and CNN networks.
   - Multi-resolution Mosaic on two examples: house and stones, apple and oranges. I get a unique, even better results than the article that I study.
   - Frist, I implemented face-swapping by overlapping two faces without warping. In whole face overlapping swapping, we can see a new artistic style is generated. Second I implemented face-swapping by using components of faces with warping, it shows that the Multi-resolution Mosaic using this CNN architecture is effective.

2 Technical approaches

2.1 Style extraction and reconstruction

These feature correlations are given by the Gram matrix $G^l = R^{N_l x N_l}$, where G^l_{ij} is the inner product between the vectorised feature map i and j in layer l:

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk}$$

To generate a texture that matches the style of a given image, we need to use a texture that matches the style of a given image. We use gradient descent from a white noise image to find another image that matches the style representation of the original image. This is done by minimizing the mean-squared distance between the entries of the Gram matrix from the original image and the Gram matrix of the image to be generated. So let $\vec{a}$ and $\vec{x}$ be the original image and the image that is generated and A^l and G^l their respective style representations in layer l. The contribution of that layer to the total loss is then

$$E_l = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} (G_{ij} - A_{ij})^2$$

And the total loss is

$$L_{style}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_l E_l$$

where wl are weighting factors of the contribution of each layer to the total loss (see below for specific values of wl in my results). The derivative of El with respect to the activations in layer l can be computed analytically:

$$\frac{\delta E_l}{\delta F^l_{ij}} = \begin{cases} \frac{1}{4N_l^2 M_l^2} ((F^l)^T (G^l - A^l))_{ji} & \text{if } F^l_{ij} > 0 \\ 0 & \text{if } F^l_{ij} < 0 \end{cases}$$

2.2 Style transfer and artistic image synthesis

To achieve artistic image synthesis, we not only need to extract style and be able to reconstruct stylistic images, if we want to apply a style to a photograph, we need to reconstruct the content of images. To reconstruct the content, we perform gradient descent on a white noise image to find another image that matches the feature responses of the original image. Using feature responses from different layers will generate different results because different layers contain different levels of features of original images. So let $\vec{p}$ and $\vec{x}$ be the original image and the image that is generated and P^l and F^l their respective feature representation in layer l. We then define the squared-error loss between the two feature representations

$$L_{content}(\vec{p}, \vec{x}, l) = \frac{1}{2} \sum_{i,j} (F^l_{ij} - P^l_{i,j})^2$$

The derivative of this loss with respect to the activations in layer l equals

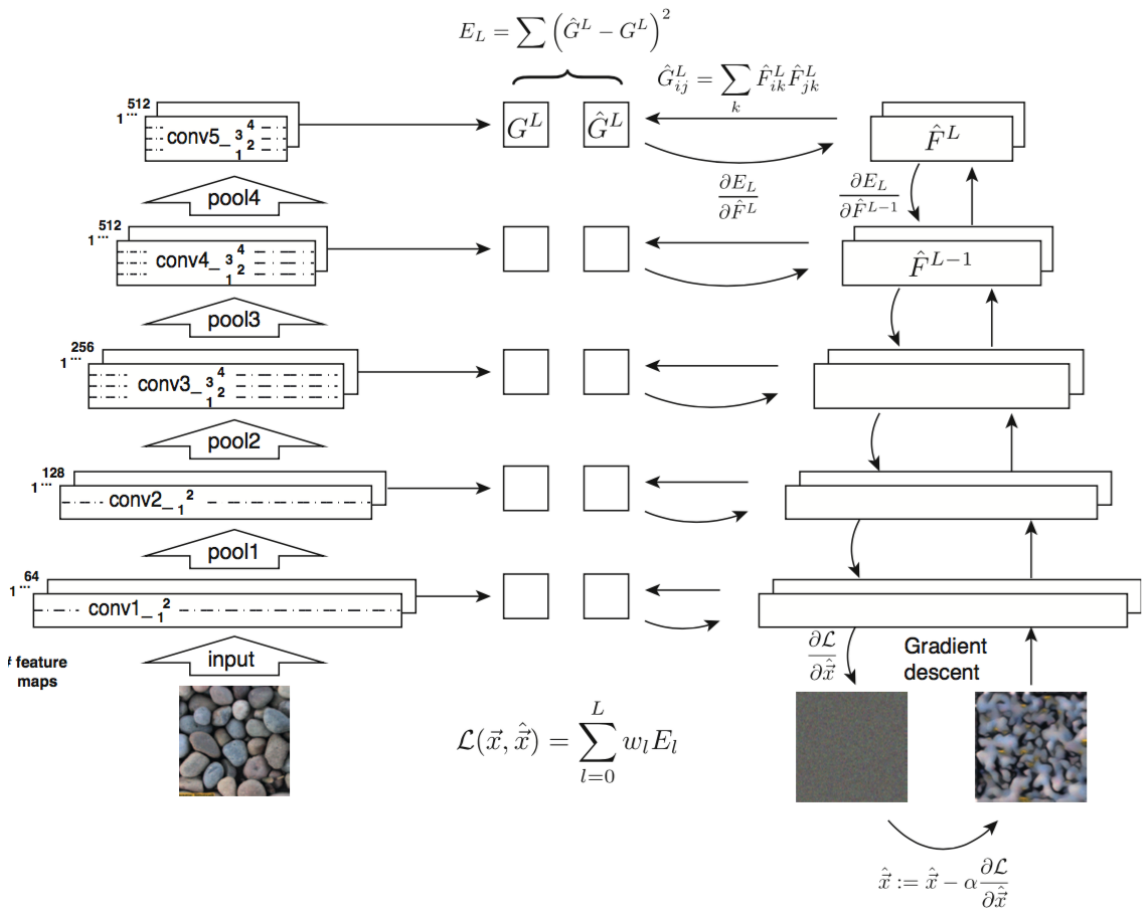


Figure 1: Style reconstruction diagram

$$\frac{\delta L_{content}}{\delta F^l_{ij}} = \begin{cases} (F^l - P^l)_{ij} & \text{if } F^l_{ij} > 0 \\ 0 & \text{if } F^l_{ij} < 0 \end{cases}$$

from which the gradient with respect to the image $\vec{x}$ can be computed using standard error back-propagation. Thus we can change the initially random image $\vec{x}$ until it generates the same response in a certain layer of the CNN as the original image $\vec{p}$.

Because different layers contain information of different levels of features, so if we use higher layer, we will generate images has higher level features but not single pixels. The resulting image will not be very high-resolution compared with the original image. Based on this, I think we can use this to achieve multi-resolution smoothing and using this to combine objects from different images.

To generate the images that mix the content of a photograph with the style of a painting we jointly minimize the distance of a white noise image from the content representation of the

photograph in one layer of the network and the style representation of the painting in a number of layers of the CNN. So let $\vec{p}$ be the photograph and $\vec{a}$ be the artwork. The loss function we minimize is

Let $\vec{p}$ be the photograph, $\vec{a}$ be the artwork. The loss function we minimise is

$$L_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha L_{content}(\vec{p}, \vec{x}) + \beta L_{style}(\vec{a}, \vec{x})$$

where α and β are the weighting factors for content and style reconstruction respectively.

By using different ratio here, we will end up with results with different effects, closer to style or closer to the original image. By adding other L_{style} term, we may have a resulting image with two styles.

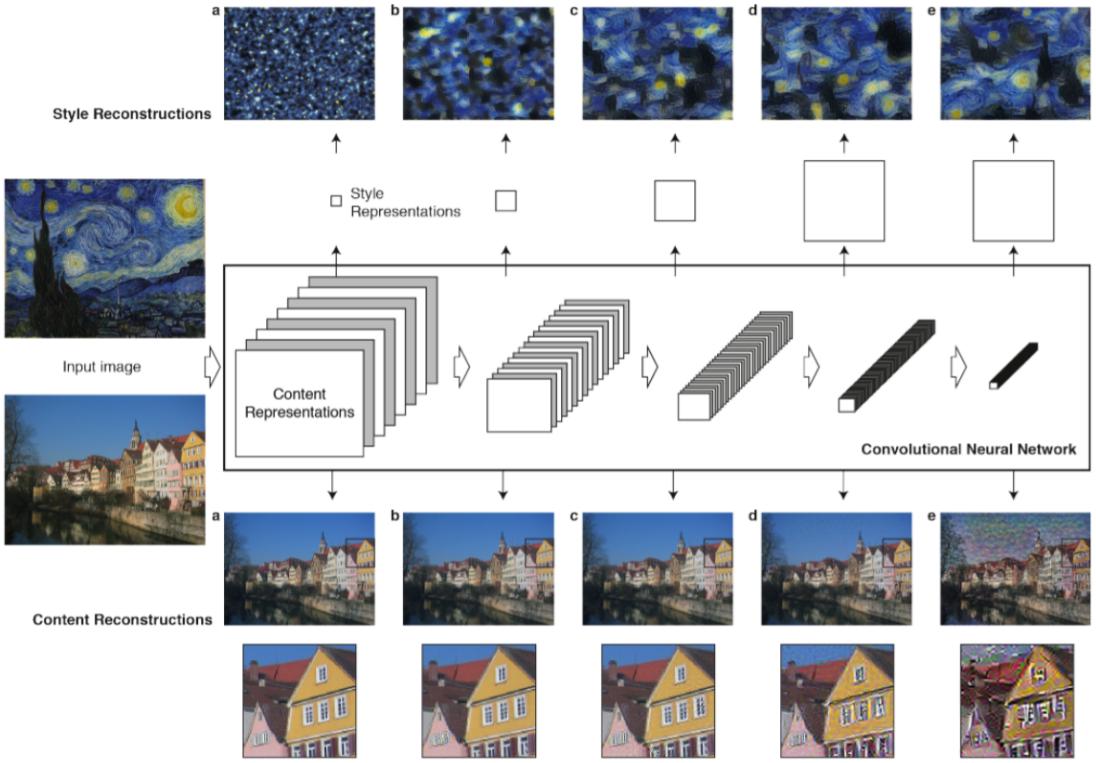


Figure 2: Artistic images synthesis diagram.

2.3 Image mosaics using a multiresolution spline

Because this CNN architecture is able to implement multi-resolution image content reconstruction. So I think it is possible to use it to do image mosaic. Here is an article discussing using a multiresolution spline to achieve Image mosaics:"A Multiresolution Spline With Application to Image Mosaics". Below are steps for achieving spline in this article:

Step 1. Laplacian pyramids LA and LB are constructed for images A and B respectively.
Step 2. Laplacian pyramids LA and LB are constructed for images A and B respectively. A third Laplacian pyramid LS is constructed by copying nodes from the left half of LA to the corresponding nodes of LS, and nodes in the right half of LB to the right half of LS. Nodes along the center line of LS are set equal to the average of corresponding LA and LB nodes.

The center line for level l of a Laplacian pyramid is at $i = 2^{N-1}$. Thus, for all i, j, l,

$$LS_l(i,j) = \begin{cases} LA_l(i,j) & \text{if } 1 < 2^{N-1} \\ (LA_l(i,j) + LB_l(i,j))/2 & \text{if } i = 2^{N-1} \\ LB_l(i,j) & \text{if } i > 2^{N-1} \end{cases}$$

Step 3. The splined image S is obtained by expanding and summing the levels of LS. The Figure3 shows the results using this multi-resolution spline to do image mosaic.

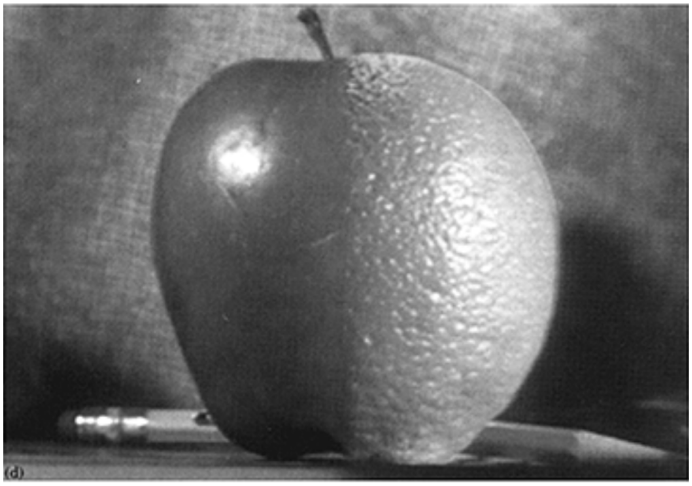


Figure 3: Image mosaics of an apple image and a orange iimage using a multiresolution spline.

Figure 4 shows the weighting functions used in this image mosaic technique. Figure 5 shows the pyramid structure for doing sampling in this image mosaic technique.

So this multi-resolution spline method performs some operations close to CNN. First, it uses weighting functions to process images, which is one type of convolution operation. Second, it uses pyramid structure in doing sampling. CNN also contains many pooling layers, so CNN also has a pyramid structure. So based on this, I think the CNN architecture used content reconstruction is able to do image mosaic as well. I do some experiments and results for image mosaic are better and more interesting than results from this article.

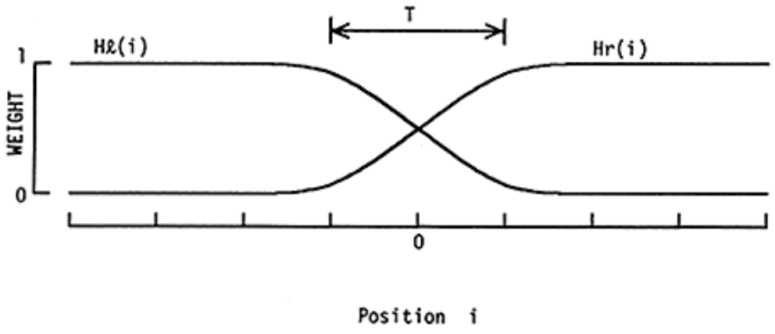


Figure 4: The weighted average method may be used to avoid seams when mosaics are constructed from overlapped images. Each image is multiplied by a weighting function which decreases monotonically across its border; the resulting images are then summed to form the mosaic. Example weighting functions are shown here in one dimension. The width of the transition zone T is a critical parameter for this method.

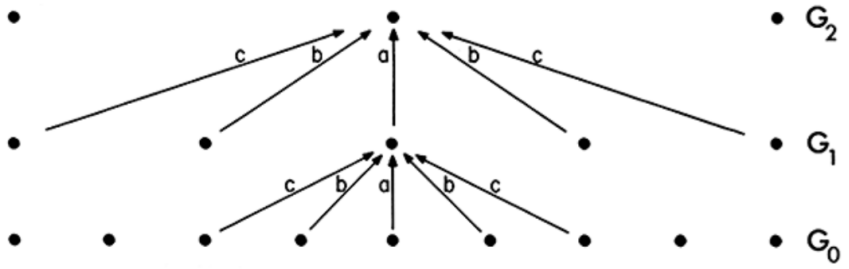


Figure 5: A one-dimensional graphical representation of the iterative REDUCE operation used in pyramid construction.

3 Results

3.1 Style reconstruction and artistic image synthesis

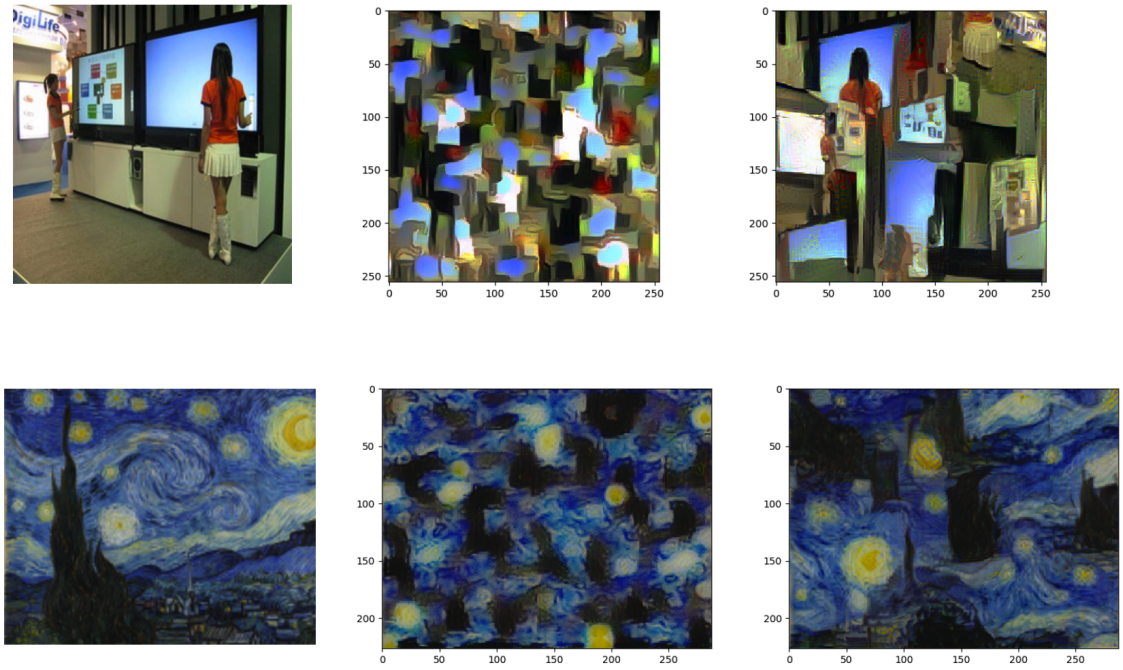
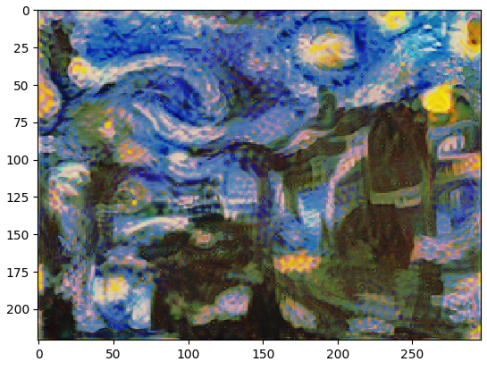


Figure 6: Style reconstruction, the left row is the original images. Right two rows are results from style synthesis, the middle row comes from a training on pool1 layer, a low layer in VGG-19, the features captured by this layer are of small size, so resulting images have features which are of small size. The right row comes from a training on pool1, pool2, pool3, pool4 4 layers, each has same weights in distance function, because we use a high layer which captures features in bigger size, so we can see more complete and bigger features, like people, screen in top image or moon and trees in bottom image.

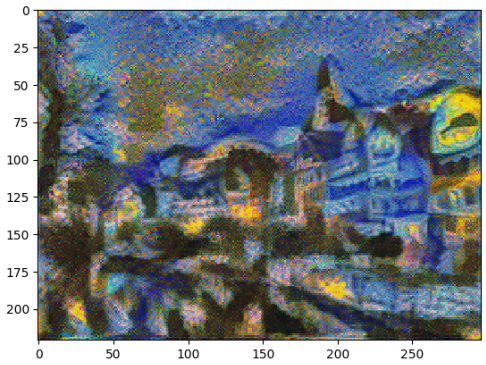
(a) The house image that is used as content

(b) Left are famous artworks that I use to extract styles as Gram matrixes, then I use the styles from these artworks and content from the house image to construct artistic images. The right row shows the result after optimizations. The ratio I use in the program for content/style is 1e-8, here because there is no normalization, so the normalization is not well defined, but with different ratios, I get images with a different ratio in style and content. The layers for content construction is *conv*4_2, because we want to use high-level features, but not low-level features like single pixels, which will make high-level features express styles. The layers for style construction is *conv*2_1, *conv*3_1, *conv*4_1, *conv*5_1.

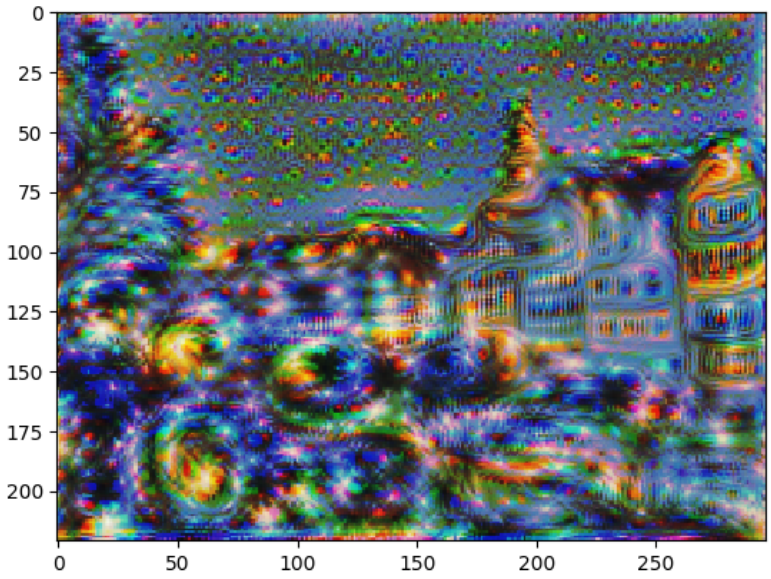
Figure 7: Artistic images synthesis



(a) When there are not enough sweeps, the content cannot show up clearly.



(b) This image has a different ratio in weights between content and style, so it has a different feeling of texture.



(c) This image uses two styles, "The Starry Night" and "The Scream". Because I use two styles as targets in different layers, it can capture the shape-spiral from stars and colors from stream-orange. The method to do it is that instead of using only one style term in loss function. I use two terms for two styles: $L_{total}(\vec{p}, \vec{a}, \vec{b}, \vec{x}) = \alpha L_{content}(\vec{p}, \vec{x}) + \beta L_{style_1}(\vec{a}, \vec{x}) + \gamma L_{style_2}(\vec{b}, \vec{x})$.

Figure 8: Some observations and further explorations in artistic images synthesis

3.2 Image reconstruction with two contents

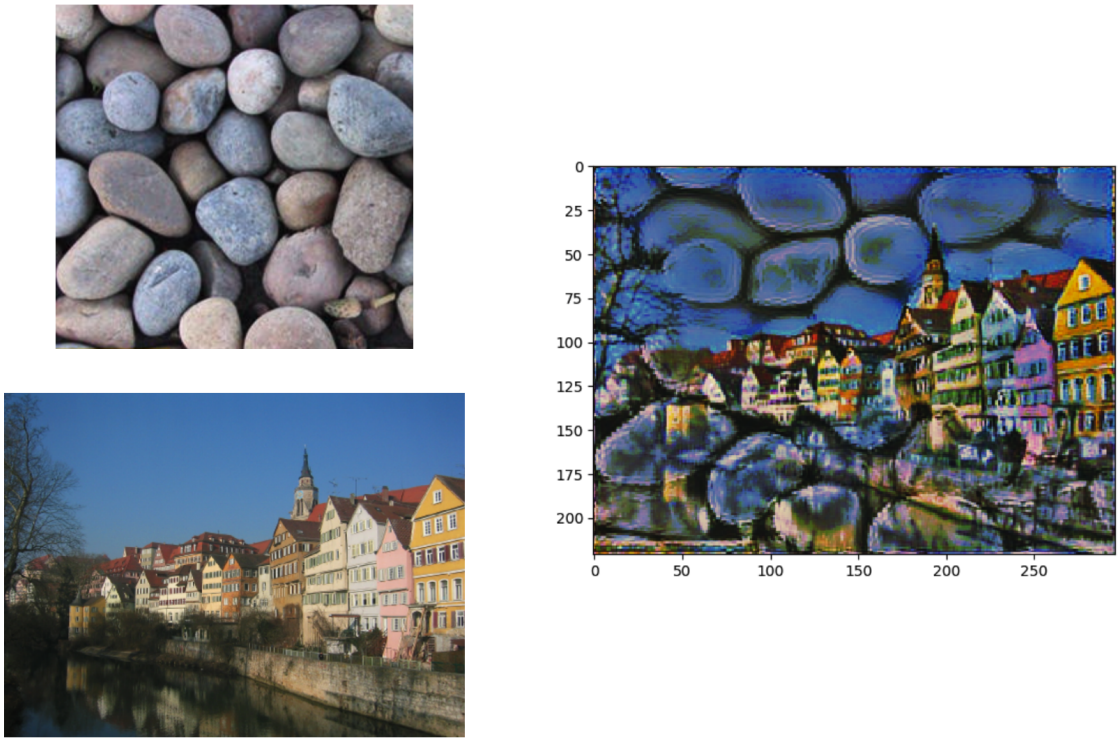


Figure 9: Left rows are two images that I extract contents from. Right row shows the result when I combine the content of these two images together. To achieve this, the loss function is defined as $L_{total}(\vec{p}, \vec{a}, \vec{x}) = \alpha L_{content_1}(\vec{p}, \vec{x}) + \beta L_{content_2}(\vec{p}, \vec{x})$, by using features in different layers, we have results where two contents appear in different ways. Different ratio of weights also yields different weights for two contents in resulting image. Here for the house image, I use $conv2\_2$ layer, for pebble image I use $conv2\_1$ layer.

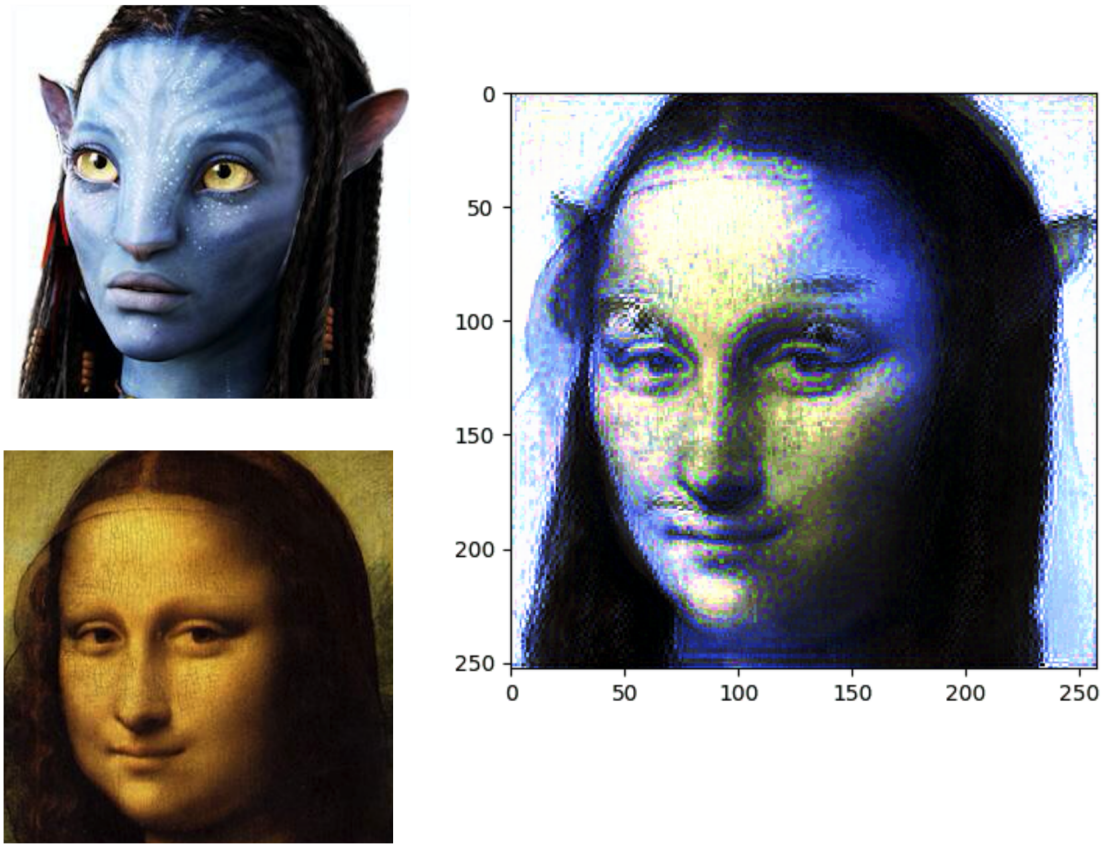


Figure 10: Left: Two famous faces that I want to combine. Right: The resulting face by using content reconstruction on these two images. Here the content from Avatar image uses $conv2\_1$ layer, the content of Mona Lisa image uses $conv2\_2$ layer. We can see that the resulting content is mainly Mona Lisa, might because Mona Lisa content uses a higher layer which has more important proportion in loss function. While we can see that due to the Avatar content, there are major changes in terms of the background, shadows, and colors. Here I think it is a unique style, which is different from previous style transfer using Gram matrix.

3.3 Multi-resolution Mosaic

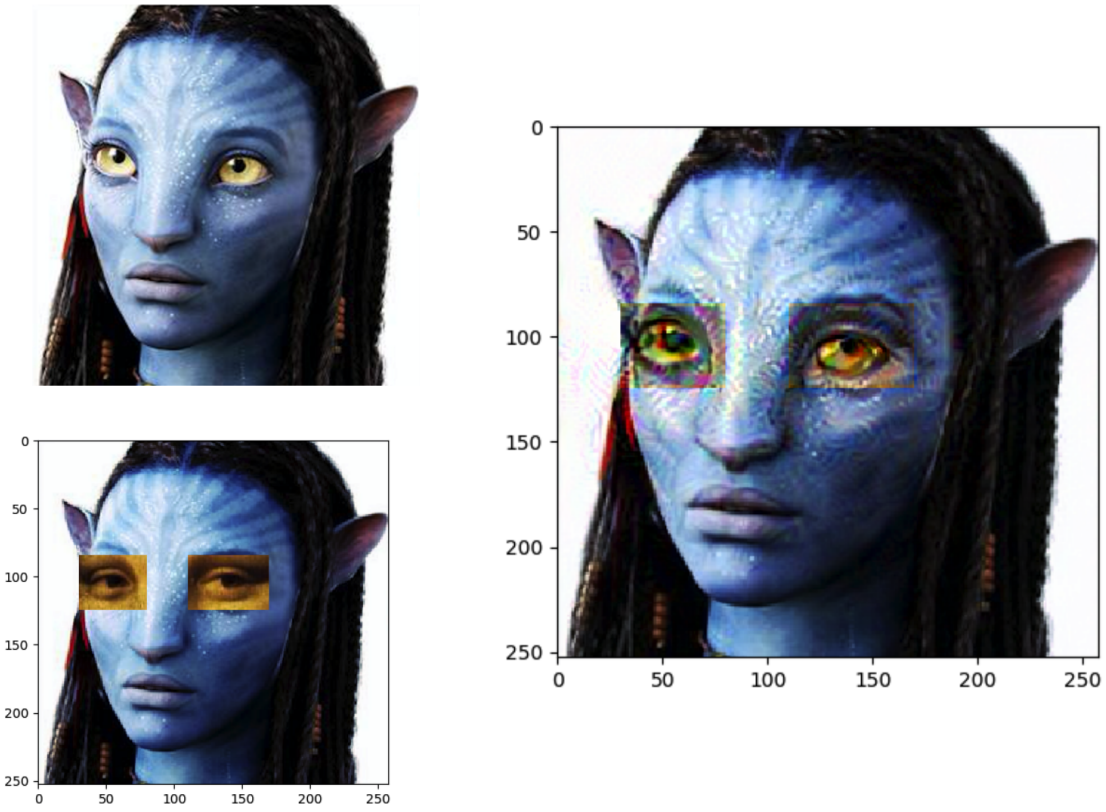


Figure 11: Left: The top is Avatar's face, the bottom is the new face that I swap the eyes with Mona Lisa's eyes, this is the image that I use as the initial input image in CNN content reconstruction optimization instead of a random noise image. Right: The resulting image by using the content feature from $conv5\_1$ layer to do content reconstruction. **Here my idea is that, by using the image from the raw combination as the initial image, after optimization by CNN, the edge between these two images should become smooth, because CNN tries to generate features on the edge that can capture target features on both sides. The loss function defined as distances between target features and resulting features controls the direction of optimization. To lower the loss function, resulting images should have features on the edge that can match features on each side.** So here we can see that between the new eyes from Mona Lisa and Avatar's face, CNN learns the local features like patterns on face well and these features match both sides pretty well. The resulting image looks natural so this new method for image mosaic is effective.

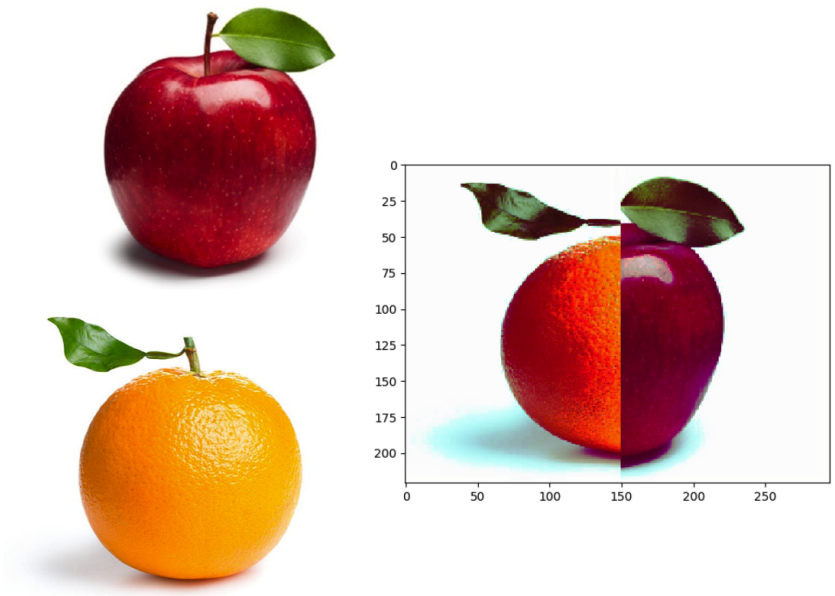


Figure 12: Left: Apple and Oranges. Right: Raw combinatin - combination of these two images before implementing image mosaic,

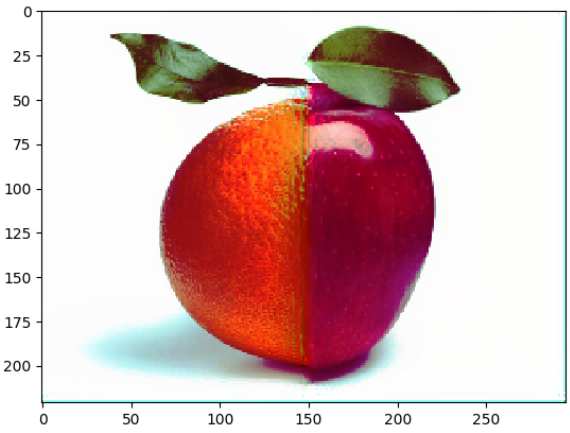


Figure 13: The resulting image by using features from *conv*2_1 layer to do content reconstruction.

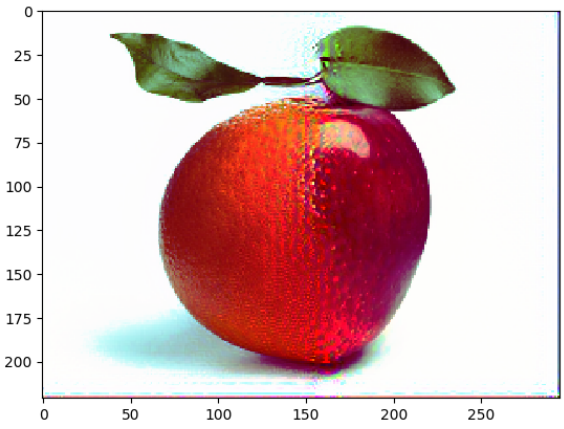


Figure 14: The resulting image by using the content feature from *conv*3_1 layer to do content reconstruction. Here we can see when using high layer which contains high-level features to do image mosaic, the resulting image has smoother edges while with lower resolution

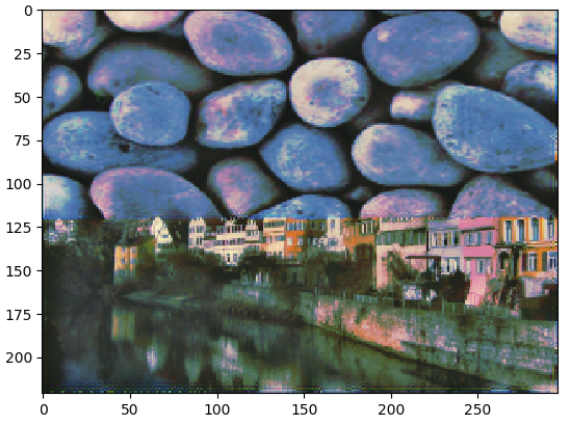


Figure 15: The resulting image by using features from *conv*1_1 layer to do content reconstruction. Which is very close to the original data layer, so the feature is very small in size, then the smoothing by matching features can not give a smooth edge.

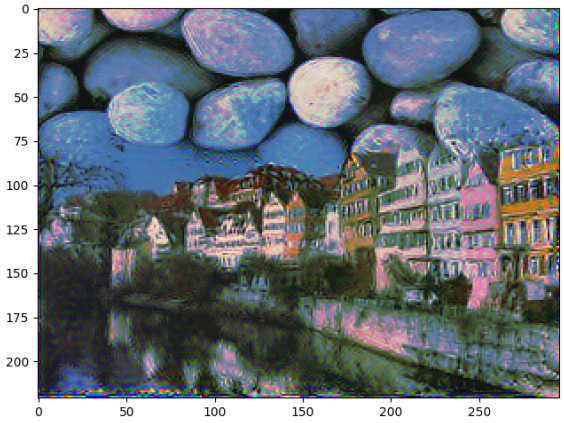


Figure 16: The resulting image by using features from *conv*3_1 layer to do content reconstruction. Because the feature is high-level features, we can see a better smoothing. Here the difference is I use random noise image as the initial image, the raw combination image - the Figure 15 on the left as the target content to do content reconstruction, so the resulting image has low resolution, this is my first method.

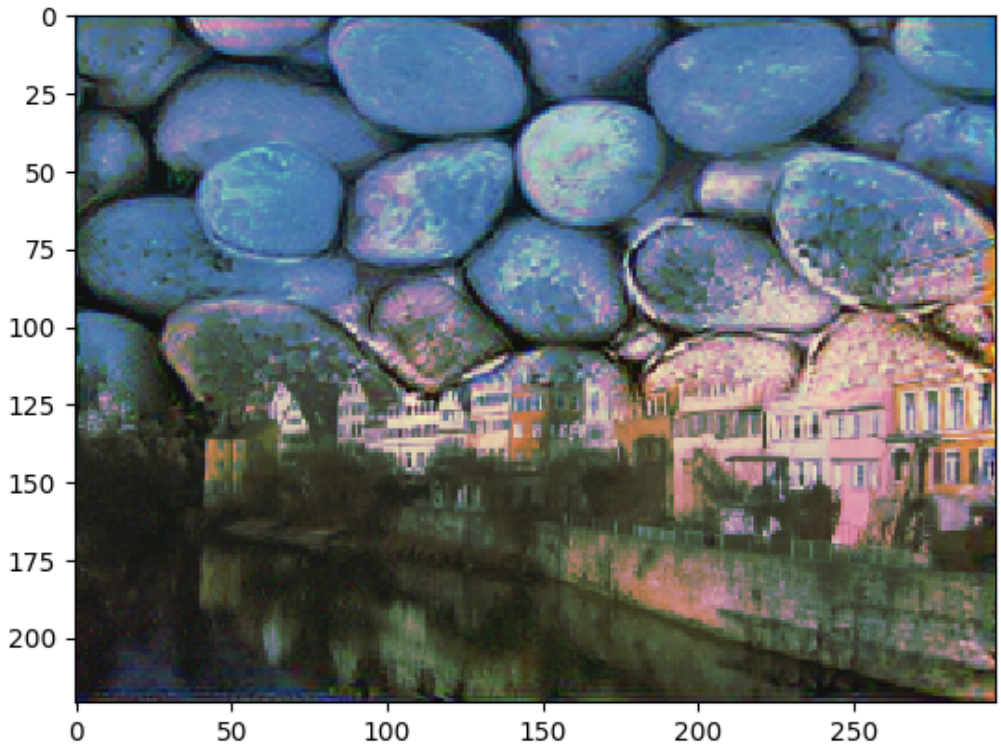


Figure 17: The resulting image by using features from *conv*3_1 layer to do content reconstruction. Here the new development is that I use the raw combination image itself as the initial image as well as the target content. By doing this, at places which are far away from the edge, the image does not change, because it is already from the target image, it contains all levels of features. While at edges, because it transfers from one image to another image, from different objects to different objects, so the high-level feature information is lost or not complete, CNN needs to optimize the resulting image to complete high-level features at edges to lower loss function which is defined as distances between features. So that's the reason why we can use the target image itself as initial image and it can help improve the resolution of resulting image.

4 Discussion

From these many experiments, it is very clear that different layer has different effects in terms of content reconstruction and image synthesis. So it still requires a lot of experiments to better understand this. Doing this also helps us to better understand CNN network. During these lots of experiments, I have some interesting observations.

- Interesting observations:
  - When doing content synthesis, $conv2\_1$ and $conv2\_2$ layer is very different. Even though they are close in CNN networks, but $conv2\_2$ is one layer higher than $conv2\_1$. Different arrangement of these two layers for contents will give very different results.
  - When doing content synthesis, when the resulting image is already close to the target image, more sweeps does not give better images in terms of appreciation. Further optimization seems even change the texture of resulting image, makes it more coarse-grained.
- Development
  - During my experiments, I already make some movies to help me understand the evolution during optimization. In the future, I will try to develop code that can provide useful functionalities such as pausing, resuming and storing several intermediate states during optimization. Thus I do not have to restart every time I find something wrong or want to slightly change some parameters. Also, it will give new ideas in doing image synthesis by providing more opportunities in optimizations.

5 My work in related code

In this project, I develop code based on the code already provided by the article: "Texture Synthesis Using Convolutional Neural Networks". But this code only has texture synthesis part, which is same as style reconstruction. But to implement Artistic image synthesis, I need to further develop the code. Here I mainly have done these designs in terms of coding:

- Code for Artistic image synthesis with one or more styles.
- Besides the original output part, I finished code for generating animations of optimization, which helps to analyze and understand the optimization process.
- Code for implementing multi-resolution mosaic.

6 References

- **Texture Synthesis Using Convolutional Neural Networks**
  Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge.
- **A Neural Algorithm of Artistic Style**
  Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge.
- **A Multiresolution Spline With Application to Image Mosaics**
  PETER J. BURT and EDWARD H. ADELSON RCA David Sarnoff Research Center