EE8601 Project 2: Image Style Transfer

Implementation of *Image Style Transfer Using Convolutional Neural Networks* by Gatys et al

Alegre, Edwin Philippe, EIT
Department of Electrical, Computer, and Biomedical Engineering
Ryerson University
Toronto, Ontario, Canada
edwin.p.alegre@ryerson.ca — 500552920

***Abstract*—The aim of this report is to document the procedures, as well as the theory gained from the implementation of Gatys' paper - *Image Style Transfer Using Convolutional Neural Networks*. This paper outlines the motivation, a brief top level theoretical outline on the workings of the above mentioned paper, the implementation (including dependencies, code, etc), as well as outcomes related to the directed studies course, EE8601 - Fundamentals of Computer Vision and Deep Learning.**

***Index Terms*—image processing, convolutional neural networks, style transfer learning**

I. Introduction

The visual arts has been a medium that conveys ones sense of wonder and abstract viewpoint of the universe to its patrons. From Vincent Van Gogh, to Picasso, to the Group of Seven, an artist wields their brush in their own manner, showcasing their own personal genius in the use of colours, strokes, and realism to convey how they see a certain subject. It is no wonder that certain artists have become synonymous with their own personal style. Van Gogh is known for his vibrant colours ad swirled strokes that are evident in his great works, with "Starry Night" as the perfect example. Picasso brings out the abstract use of shapes in his work. Monet captures the vibrant realism of his landscape masterpieces. Artists have tried to copy that style and take years to reach a level that was achieved by these masters. Recent technological advancements have allowed for the use of *convolutional neural networks* (CNN) to achieve great feats in the field of computer vision and engineering. An example of image style transfer can be seen in Figure 1.

This paper will focus on the implementation of *Image Style Transfer Using Convolutional Neural Networks* by Gatys et al. This paper utilizes convolutional neural networks to extract the *style* of one image (the way an artist paints) to an image with a different *content*. In the world of computer vision this is known as *style transfer*. This is done with the use of CNNs that extract the style of one image and combines it with another.

Figure 1. An example of image style transfer. The top left photo showcases Stanford University's Tower and is known as the *content image*. The top right photo is that of Van Gogh's Starry Night and is known as the *style image*. By applying the image transfer technique, a synthesized image of the two input images is formed and shown at the bottom

II. Dependencies

In order to successfully implement this project, the following technical dependencies are needed.

A. Python (ver. 3.7.7, 64-bit)

Python is the primary prototyping language used for machine learning, computer vision, and image processing applications. As such, it is the programming language used to implement this project.

It is important to note that Python will be supplemented by Tensorflow and Keras (both tools that correspond with easily implementing machine learning models within the Python environment).

It is highly recommended that Python is supplemented with *pip*, a Python package manager that makes it easier to install libraries to use within Python. An ideal IDE that the reader is comfortable with using is also suggested.

B. Tensorflow 2.0

Tensorflow is an open source machine learning library, created by the Google Brain Team, that is capable of interfacing with Python. Tensorflow is widely used in academia and the industry, which makes it a viable platform to use for this project. Specifically, this model relies on a CNN model known as VGG19 which needs to have been trained using a specific data set known as *ImageNet*. Luckily, Tensorflow has already implemented this exact model and set for use.

Tensorflow is capable of training and running deep learning models. The most recommended API to use when dealing with neural networks on Tensorflow is called Keras. It is worth noting that Tensorflow may be used locally (on the host machine) or with a cloud computing service such as Amazon Web Services (AWS) in order to run the models utilized. While it is entirely possible to run Tensorflow locally, it may be time consuming as the vanilla Tensorflow package utilizes the host CPU in order to process the model. This method takes approximately 1 to 2 hours to run the code. The time used to run the model can be optimized by the use of cloud computing or though outsourcing the processing to a local machine's GPU. Typically, GPU outsourcing is achieved through the use of CUDA.

C. CUDA (optional)

CUDA is a parallel computing platform created by NVIDIA that enables the use of a local machine's GPU for general purpose computing. Specifically, the use of CUDA in this project allows our model to return the results at a much quicker rate due to the use of the local machine's dedicated GPU instead of its CPU. CUDA relies on a variety of dependencies in order to interface with Tensorflow. The following dependencies must be met before CUDA can be utilized:

- NVIDIA GPU card with CUDA Compute Capability of 3.5 or higher
- NVIDIA GPU Drivers - CUDA 10.1 requires 418.x or higher
- CUDA Toolkit - Tensorflow 2.1.0 requires CUDA 10.1
- CUPTI - Installed along with the CUDA Toolkit
- cuDNN SDK

D. Google Colab (optional)

Google Colab is a notebook service that allows for the creation and execution of Python code within the Google interface. By uploading Python code to a linked Google Drive, code can be run online. Furthermore, Google Colab allows for free GPU utilization, which is extremely useful if CUDA is not a viable solution due to the hardware constraints of the local machine.

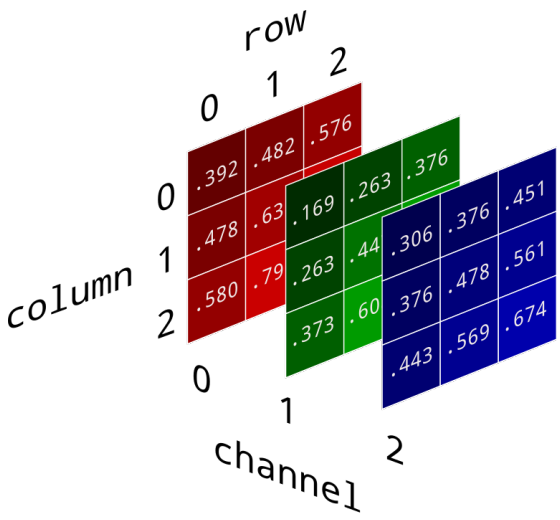

Figure 2. Visual representation of an RGB image

III. Background and Related Work

In order to properly understand how Gatys' algorithm works, a background in various concepts in Image Processing and Deep Learning is needed. The main concepts are discussed below.

A. Image Representation

In the realm of engineering, images are defined as a matrix of pixel values of size MxN. Typically, the pixel intensity is represented by 1 *byte* with a range of [0, 255]. Greyscale images are restricted to this range and are represented by a single matrix. Color images are typically represented as a 3 channel matrix that corresponds to red, green, and blue, which is why colour images are known as RGB images of size MxNx3. An example of an RGB image is show in Figure 2.

B. The Convolution Operation

The *Convolution* Operation is the backbone on which Convolutional Neural Networks are built on. Convolution is an elementary operation in the field of electrical engineering, specifically in signal processing and image processing. Convolution works by sliding a *filter* or a *kernel* of a pre-specified dimension (typically 3x3) over a matrix. Each element in the matrix is multiplied by the overlaying kernel value and all of the products are summed to produce a single output. Typically, convolution involves flipping the matrix, especially in the field of signal processing. However, in the case of computer vision and CNNs, the matrix remains unflipped. This specific case of convolution is known as *cross correlation*

Specifically, convolution in CNNs is used to extract features from the input image. By using small filters of input data, convolution maintains the relationship between pixels by

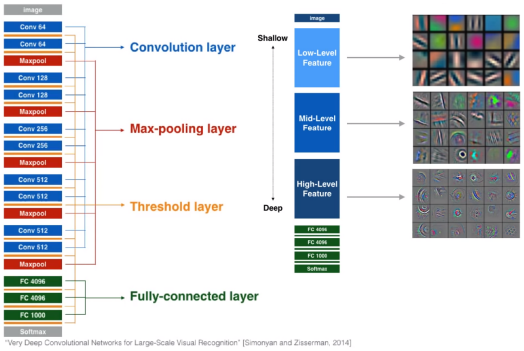

Figure 3. A generic CNN based off the VGG19 network that shows all the typical layers used. The smaller figure on the right side displays what types of features are captured at each convolutional layer in the CNN

identifying image features. The resultant matrix formed by the convolution within a convolutional layer is known as a *feature map*. This feature map is a 3D matrix of size N_{width} x N_{height} x $N_{convolution}$ where N_{width} is the width of the input image, N_{height} is the height of the input image, and $N_{convolution}$ is the number of filters in the layer.

C. Zero-Padding

Zero Padding is another concept widely used in signal and image processing. Zero padding surrounds the input data by a border of zeros such that the borders of the input image can also be captured int eh feature map during the convolution. This also allows for preservation of the dimensions when the input data is convolved with a filter.

D. Convolutional Neural Network, CNN

The *Convolutional Neural Network* is broken down into a smaller set of functions of made up of layers, where each one of those layers are a neural network in of itself. Specifically, the CNN is comprised of 3 main layers, those being: *the convoluton layer*, *the pooling layer*, and *the threshold/non linear layer*. Depending on the CNN architecture, there may be more layers involved in the CNN, but these 3 layers remain constant. The CNN works extremely well for image based problems. Specifically, the architecture of the CNN allows it to learn a variety of features, from low-level features (lines, edges, etc) to mid-level features (combined lower level features that showcase patterns) high-level features (truly sophisticated and complex features/textures at the top level). The initial layers usually learn the low-level features, with each progressive layer learning higher-level features. An example of a CNN is shown in Figure 3, specifically showcasing the CNN used in this project, the VGG19 network. THe VGG19 network contains 3 fully connected layers and 1 softmax layer, which are not utilized in this project. Hence it will not be discussed, however each utilized layer is explained in more detail below.

1) The Convolution Layer: The convolution layer is the most important layer in the CNN and is the name sake of

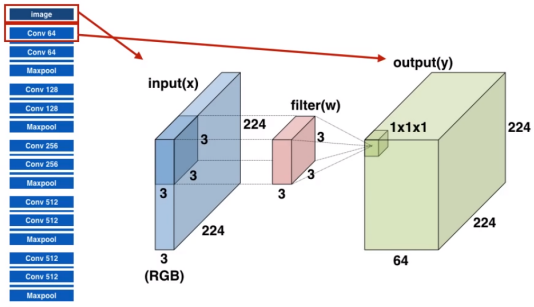

Figure 4. A graphical example of how a convolutional layer in a CNN works. Note that the actual convolutional layer is comprised of a tensor with size 224x224x64

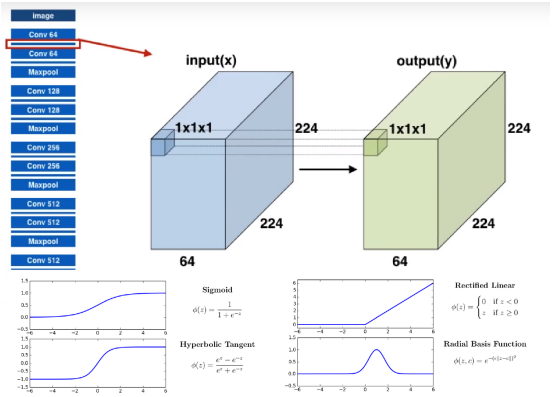

Figure 5. A graphical representation of the threshold layer in a CNN. The threshold is based off the activation function used in the network. The 4 most common activation functions are shown at the bottom of the figure

the CNN. In general with the case of images as the input, the convolution layer is a 3D convolution. The input image is convolved with a kernel where the output is the basis for the feature map. In the case of the VGG19, the first layer generates as 224x224x64 feature map, which is due to 64 different kernels being used to produce the different depths of this layer. The first stage convolution layer of the VGG19 network is shown in Figure 4.

2) The Threshold/Non Linear Layer: Following the Convolution Layer is the *threshold layer*. This layer introduces a non linear component to the network, where the most used activation function is normally the *rectified linear unit (ReLU)* but other functions such as the hyperbolic tangent can also be used. This layer is the same size as the previous feature map and each element corresponds to the same element space of the feature map before it. This can be seen in Figure 5

3) The Pooling Layer: The *pooling layer* is essentially the layer that downsamples its input layer in space. For example, the first pooling layer in the VGG19 network takes an input from the threshold layer of size 224x224x64 and applies *max pooling* onto it. Usually, pooling is done by sliding over the input layer with a 2x2 kernel of *stride* 2, meaning it moves two pixels over instead of the usual one.

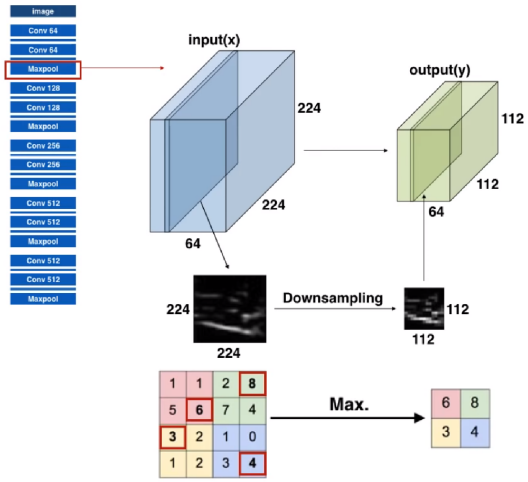

Figure 6. A graphical representation of the pooling layer in a CNN. Note that this layer essentially downsamples the input feature map. The figure at the bottom shows how max pooling is achieved, where a 2x2 kernel with a stride of 2 slides over the input matrix and translates the maximum value in the viewing window as the new downsampled pixel in the resultant pooled feature map.

This kernel performs a max function over its window, choosing the maximum element to represent that feature area in the downsampled feature map. There exists other pooling methods such as the *average pooling* or *sum pooling*. Gatys states that average pooling yields better results, which is why it is utilized as the pooling method for this implementation. As specified earlier, the first pooling layer in VGG19 takes a 224x224x64 input feature map and returns a 112x112x64 downsampled feature map. In all cases, pooling helps make representations become approximately invariant to small translations to the input. Invariance to translation causes the values of most of the pooled output values to stay the same despite the input being translated by a small amount. Invariance to translation is particularly useful if the presence of the feature is more important than the location of it. A sample of the pooling layer can be seen in Figure 6

E. VGG19 Network

The Visual Geometry Group Network (VGG) is the winning submission of the 2014 ImageNet challenge. This network has two versions, the VGG16 and VGG19 networks. The number symbolizes the amount of layers in the network, with VGG16 having 16 layers and VGG19 having 19 layers. VGG19 is used for this application and consists of 16 convolutional layers and a rectified linear unit (ReLU) as its non-linear threshold function. The convolutional layers are separated by 5 pooling layers and ends with 3 fully connected layers. As stated previously, the fully connected networks will not be used for this project. Specifically, the content representation will be matched using 'conv4_2', and the style representation will be matched using 'conv1_1', 'conv2_1', 'conv3_1', 'conv4_1', and 'conv5_1'. Each layer will have a weight of 1/5. The graphical representation of the VGG19 network can be seen in Figure 7

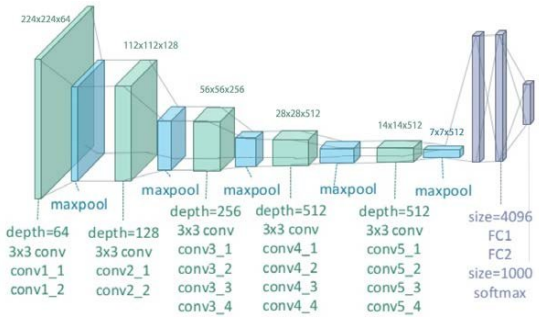

Figure 7. The VGG19 Network. Notice the last 3 layers of the network, known as the fully connected layers. While these layers are part of the VGG19 network, they will not be utilized in this implementation

IV. Deep Image Representation

The main takeaway from this paper is that **the representations of the content and style are separable and thus, both representations cna be manipulated independently to produce an output of the user's liking**. Using this finding, it was determined that the overall loss function is a linear combination of the two individual loss functions. Thus, by regulating the individual loss functions, the emphasis on the resultant image can be based on either the reconstruction of the content or style input. This will be discussed in further detail in the upcoming sections.

A. Content Representation

Given an input image $\vec{x}$, this input image is passed into each layer of the CNN, providing a feature representation at each layer, F^{l}. Gradient descent can be applied to a white noise image in order to visualize the image information that is encoded at every layer of the network. This provides another image that can be used to match the feature responses of the original image. These responses in a layer, l, can be stored in a matrix with the generalized form of $X^{l} \in \mathcal{R}^{N_l \text{x} M_l}$, where each layer contains N_l distinctive filters and thus, has N_l each of size M_l, where M_l is the height times the width of the feature map. From this, the activation of the i^{th} filter at position j in layer l can be symbolized as X^{l}_{ij}

These two images can be called the *content image*, $\vec{p}$ with feature representation (content features) $\vec{P}$ generated in Layer 4_1, and the *generated image*, $\vec{x}$ with feature representation (content features) $\vec{F}$ generated in the same layer.

Thus, the squared-error loss between the two feature representations can be defined as

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

From the derivative of this loss function, defined below, the gradient descent with respect to the generated image, $\vec{x}$, can be computed using standard back-propagation.

$$\frac{\partial \mathcal{L}_{\text{content}}}{\partial {F^{\text{l}}}_{\text{ij}}} = \begin{cases} ({F^{\text{l}} - P^{\text{l}}})_{\text{ij}}{}^2 & if {F^{\text{l}}}_{\text{ij}} > 0 \\ 0 & if {F^{\text{l}}}_{\text{ij}} < 0 \end{cases}$$

As stated in earlier sections, CNNs are invariant to translation but are increasingly sensitive to the actual *content* of the image. To reiterate, the lower layers capture the pixel based relationships of the features, such as line and edges, whereas the higher layers capture the actual content of the image. It is for this reason that the feature responses of the higher layers are referred to as the *content representation*.

B. Style Representation

To generate the *style representation* of an image, a feature space, which can be built on top of the filter responses in any layer, is used to obtain the texture information from this image. Specifically, this feature space contains the feature correlations between different filter responses. These correlations are computed through the *Gram matrix*, $\mathcal{G}^{\text{l}} \in \mathcal{R}^{\text{N}_\text{l} \text{ x } \text{N}_\text{l}}$, where ${\mathcal{G}^{\text{l}}}_{\text{ij}}$ is the inner porduct between the vectorised feature maps i and j in the layer l:

$${\mathcal{G}^{\text{l}}}_{\text{ij}} = \sum_k {F_{\text{ik}}}^{\text{l}} {F_{\text{jk}}}^{\text{l}}$$

Just like the content representation, this gradient descent can be used on a random white noise image to minimise the mean-squared error to achieve a similar visualization. However, this optimization utilizes the Gram matrices from the original image $\vec{a}$ and the generated image $\vec{x}$. The respective style representations in layer l would then be A^{l} and G^{l}. The contribution of layer l to the total loss would then be

$$E_{\text{l}} = \frac{1}{4{N_{\text{l}}}^2 {M_{\text{l}}}^2} \sum_{i,j} ({G_{\text{ij}}}^{\text{l}} - {A_{\text{ij}}}^{\text{l}})^2$$

Thus, the total style loss would then be:

$$\mathcal{L}_{\text{style}}(\vec{a}, \vec{x}) = \sum_{l=0}^{L} w_{\text{l}} E_{\text{l}}$$

where w_{l} are the weights of the contribution of each layer to the total loss. Since there are 5 layers used for the style transfer algorithm, each layer is weighted at $w_{\text{l}} = \frac{1}{5}$.

Once again, similar to content representation, the derivative of this mean-squared error,defined below, can be taken with respect to the activation in layer l. From this, the gradient of E_l with respect to the generated image $\vec{x}$ can be computer using standard back-propagation

$$\frac{\partial E_{\text{l}}}{\partial {F^{\text{l}}}_{\text{ij}}} = \begin{cases} \frac{1}{{N_{\text{l}}}^2 {M_{\text{l}}}^2} ((F^{\text{l}})^{\text{T}} (G^{\text{l}} - A^{\text{l}}))_{\text{ij}} & if {F^{\text{l}}}_{\text{ij}} > 0 \\ 0 & if {F^{\text{l}}}_{\text{ij}} < 0 \end{cases}$$

C. Style Transfer

The concept of *style transfer* involves the synthesizing of a new image that simultaneously matches the content representation of a photograph $\vec{p}$ and the style representation of $\vec{a}$. The mathematical problem of style transfer essentially boils down to an optimization problem. Specifically, it is the joint minimization of a white noise image's, $\vec{x}$, feature representation distance from the photograph's content representation in one layer and the artwork's style representation on a number of layers. The total cost function to me minimized is:

$$\mathcal{L}_{\text{total}}(\vec{p}, \vec{a}, \vec{x}) = \alpha \mathcal{L}_{\text{content}}(\vec{p}, \vec{x}) + \beta \mathcal{L}_{\text{style}}(\vec{a}, \vec{x})$$

where α and β are weighing factors for the respective content and style representations. The gradient with respect to the generated image $\frac{\partial \mathcal{L}_{\text{total}}}{\partial \vec{x}}$ can be used for a numerical optimisation strategy. Gatys et al utilized the L-BFGS optimizer, however, this is not available in Tensorflow. Thus, the ADAM optimizer is used as a substitute. The style image is always resized to the same size as the content image before computing the feature representation in order to extract image information on comparable scales.

V. Summary of the Algorithm

The summary of the algorithm is shown in Figure 8. The main steps are as follows:

1) Provide the input images for the style $\vec{a}$ and content $\vec{p}$ and are passed through the network.
2) The style representation P^l is computed on all layers. The style Gram matrix at each layer is also calculated from this feature representation.
3) The content representation A^l is computed on one layer.
4) The randomly generated white noise image is passed through the network to generate its style features G^l and its content features F^l.
5) On each layer included in the style representation, the mean-squared error is calculated between the Gram matrix of the generated image G^l and the Gram matrix of the original image A^l to provide the style loss $\mathcal{L}_{\text{style}}$.
6) On the one layer included in the content representation, the mean-squared error is calculated between the generated content features F^l and the photograph's content representation P^l to provide the content loss $\mathcal{L}_{\text{content}}$
7) Since the total loss $\mathcal{L}_{\text{total}}$ is a linear combination of the content and style losses, use these two losses to compute it.

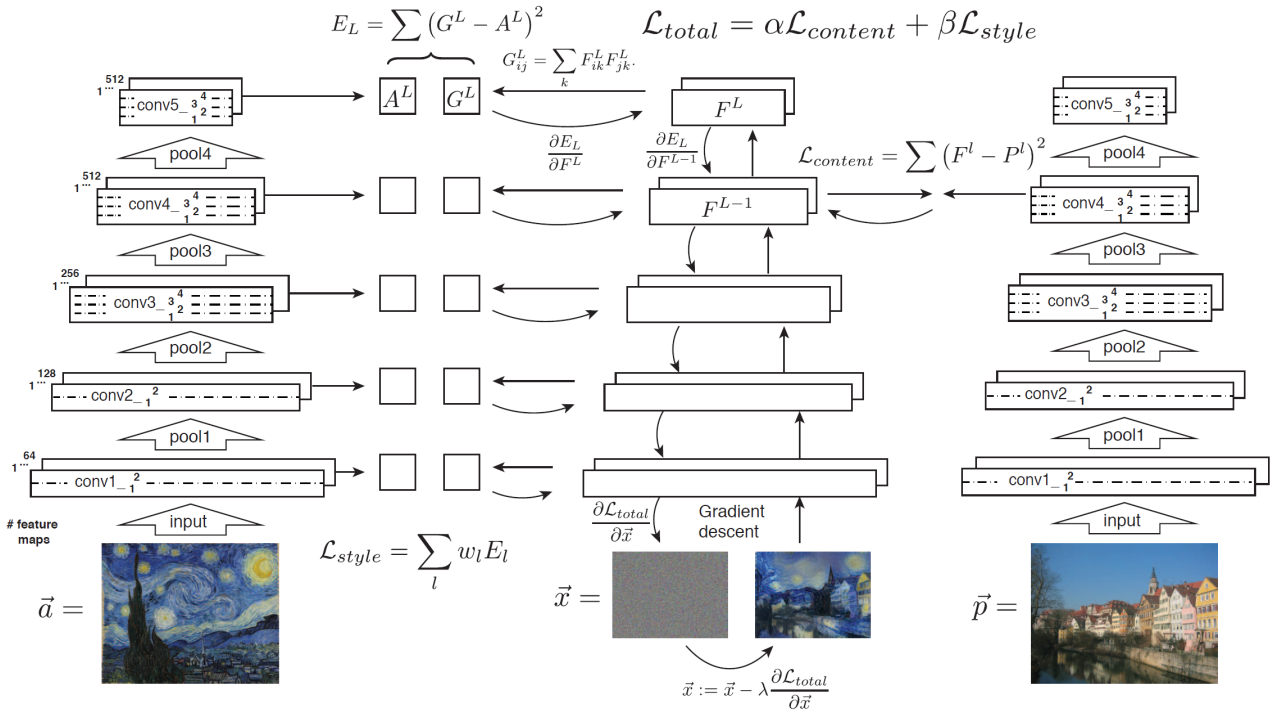

Figure 8. Style Transfer Algorithm. The detailed steps are discussed in Section V - Summary of the Algorithm

8) The derivative of the total loss with respect to the pixel values of the generated image $\frac{\partial \mathcal{L}_{\text{total}}}{\partial \vec{x}}$ is computed through standard back propagation.
9) The gradient computed in the previous step is used as an input into the ADAM optimizer and is iterated until it simultaneously matches the style features of the style image $\vec{a}$ and content photograph $\vec{p}$

VI. Python Code

The following section details the code used for this project. The first program involves functions that are utilized for the image utilities. This includes loading the image, processing the image for the VGG19 network, deprocessing it back as an image, and saving the image.

The second code is the main function. It outlines the functions used for setting up the VGG19 model, as well as the loss functions and feature extractions. The last part of the code is the actual execution that runs the training steps and saves the resulting image.

Appendix A, which is attached with this report, shows a step by step result of the code as it's run. It was executed using Google Colab and demonstrates the output and flow of how image style transfer actually works based off the written code.

A. Image Utilities

```
# -*- coding: utf-8 -*-
"""
Created on Mon Apr  6 13:26:18 2020

@author: ealegre
"""

import tensorflow as tf
import numpy as np
from PIL import Image
import matplotlib as plt

# IMAGE UTILITIES - FUNCTION 1 - LOAD IMAGE

def load_img(img_path):
    # Read the image and convert the computation graph to an image format
    img = tf.io.read_file(img_path)
    img = tf.image.decode_image(img, channels=3)
    img = tf.image.convert_image_dtype(img, tf.float32)

    # Setting the scale parameters to change the size of the image
    # Get the width and height of the image. Cast it to a float so it can be divided
    shape = tf.cast(tf.shape(img)[:-1], tf.float32)
    # Set the absolute maximum dimension for the image
    # max_dim = 1024
    max_dim = 512
    # Find which side is the longer side, this will be used to generate our scale
    max_side = max(shape)
    scale = max_dim / max_side
    new_shape = tf.cast(shape * scale, tf.int32)
    img = tf.image.resize(img, new_shape, method=tf.image.ResizeMethod.BILINEAR)

    img = img[tf.newaxis, :]
    return img

# IMAGE UTILITIES - FUNCTION 2 - DEPROCESS IMAGE
```

```
def deprocess_img(processed_img):
    processed_img = processed_img*255
    processed_img = np.array(processed_img, dtype=np.uint8)
    if np.ndim(processed_img)>3:
      assert processed_img.shape[0] == 1
      processed_img = processed_img[0]
    return Image.fromarray(processed_img)

# IMAGE UTILITIES - FUNCTION 3 - SAVE IMAGE

def save_img(best_img, path):
    img = Image.fromarray(best_img)
    img.save(path)

def imshow(image, title=None):
  if len(image.shape) > 3:
    image = tf.squeeze(image, axis=0)

  plt.imshow(image)
  if title:
    plt.title(title)
```

B. *Model Utilities & Main Function*

```
# -*- coding: utf-8 -*-
"""
Created on Mon Apr  6 13:16:38 2020

@author: ealegre
"""

# DEPENDENCIES, VARIABLES, AND PATHS

# Libraries
import tensorflow as tf
from tensorflow.python.keras import models
import os
import glob
import image_utils as IU

# Define paths
content_path = 'input/content/elon.jpg'
style_path = 'input/style/ironman.jpg'
init_path = 'input/init/512.jpg'
lena = 'input/content/lena_test.png'
lion = 'input/content/lion.jpg'
dog = 'input/content/dog.jpg'

# Load the style and content images
style_img = IU.load_img(style_path)
content_img = IU.load_img(content_path)
generated_img = tf.Variable(content_img)

# Define which layers are to be used for this model. These layers are defined in Section
# 3 of Gatys' paper
content_layers = ['block5_conv2']
style_layers = ['block1_conv1',
                'block2_conv1',
                'block3_conv1',
                'block4_conv1',
                'block5_conv1']

num_content_layers = len(content_layers)
num_style_layers = len(style_layers)

# MODEL UTILITIES - FUNCTION 1 - VGG19 LAYERS

def get_vggLayers(style_or_content_layers):
    """
    Creates the model with intermediate layer access

    This function will load in the VGG19 model used in Gatys' paper, with access to the
    intermediate layers. A new model will be generated by using these layers that will take an
    input image and return an output from the intermediate layers from the VGG19 model

    Returns
    -------
    A Keras model that takes inputs and outputs of the style and content intermediate layers

    """
    # Instantiate the VGG19 model. We are not including the fully connected layers, instantiate the weight based off ImageNet training
    vgg = tf.keras.applications.VGG19(include_top=False, weights='imagenet')

    # VGG19 model is already trained off ImageNet, set trainable to False
    vgg.trainable = False

    # All we are doing is clipping the model accrodingly. Just set the input of the model to be the same as its original
    # The outputs are going to be set to the layers specified in either the style or content layers, this will
    # be set once you've passed one of them into the function as an argument
    vgg_input = [vgg.input]
    vgg_output = [vgg.get_layer(name).output for name in style_or_content_layers]
    model = tf.keras.Model(vgg_input, vgg_output)

    return model

# MODEL UTILITIES - FUNCTION 2 - THE GRAM MATRIX

def gram_matrix(input_tensor):
  """
  Generates the Gram matrix representation of the style features. This function takes an input tensor and
  will apply the proper steps to generate the Gram matrix

  Returns
  -------
  A tensor object Gram matrix representation

  """
  # Equation (3)

  # Generate image channels. If the input tensor is a 3D array of size Nh x Nw x Nc, reshape
  # it to a 2D array of Nc x (Nh*Nw). The shape[-1] takes the last element in the shape
  # characterstic, that being the number of channels. This will be our second dimension
  channels = int(input_tensor.shape[-1])

  # Reshape the tensor into a 2D matrix to prepare for Gram matrix calculation by multiplying
  # all of the dimensions except the last one (which is what the -1 represents) together
  Fl_ik = tf.reshape(input_tensor, [-1, channels])

  # Transpose the new 2D matrix
  Fl_jk = tf.transpose(Fl_ik)

  # Find all the elements in the new array (Nw*Nh)
  n = tf.shape(Fl_ik)[0]

  # Perform the Gram matrix calculation
  gram = tf.matmul(Fl_jk, Fl_ik)/tf.cast(n, tf.float32)

  # Generate the Gram matrix as a tensor for use in our model
  gram_tensor = gram[tf.newaxis, :]
  return gram_tensor

# MODEL UTILITIES - FUNCTION 3 - STYLE & CONTENT MODEL CLASS

class StyleContentModel(tf.keras.models.Model):
  def __init__(self, style_layers, content_layers):
    super(StyleContentModel, self).__init__()
    self.vgg = get_vggLayers(style_layers + content_layers)
    self.style_layers = style_layers
```

```
    self.content_layers = content_layers
    self.num_style_layers = len(style_layers)
    self.vgg.trainable = False

  def call(self, inputs):
    # Expects a float input between [0, 1]
    inputs = inputs * 255.0
    preprocessed_input = tf.keras.applications.vgg19.preprocess_input(inputs)
    outputs = self.vgg(preprocessed_input)
    style_outputs, content_outputs = (outputs[:self.num_style_layers], outputs[self.num_style_layers:])

    style_outputs = [gram_matrix(style_output) for style_output in style_outputs]

    content_dict = {content_name:value for content_name, value in zip(self.content_layers, content_outputs)}

    style_dict = {style_name:value for style_name, value in zip(self.style_layers, style_outputs)}

    return{'content':content_dict, 'style':style_dict}


feature_extractor = StyleContentModel(style_layers, content_layers)

# Load the style and content images
style_img = IU.load_img(style_path)
content_img = IU.load_img(content_path)

# Extract the features from the style and content images
style_features = feature_extractor(style_img)['style']
content_features = feature_extractor(content_img)['content']

# The Tensorflow variable function initializes our image to be used for gradient descent.
# This image has to be the same size and type as the content image, which means that as long as it's loaded in before being
# called upon as the generated image, it should be fine to use. This is essentially the image that will be optimized.
# In case you want to use another image, you can uncomment the code below
'''
init_img = load_img(content_path)
generated_image = tf.Variable(init_img)
'''
generated_img = tf.Variable(content_img)

# Remember that all of our images are read in as a float32 type, so we have to define the range as [0, 1] to keep it within 255
def clip_range(img):
  return tf.clip_by_value(img, clip_value_min=0.0, clip_value_max=1.0)

# Choose an optimizer. The paper chose L-BFGS but Tensorflow doesn't have that, so we'll use ADAM
optimizer = tf.optimizers.Adam(learning_rate=0.02, beta_1=0.99, epsilon=1e-8)

# Remember that we are trying to optimize the Total Loss function. However, there are values that correspond to the style and content weight
# The Alpha value corresponds to content weight and the Beta corresponds to the style weight. Let's define these
alpha = 1e4
beta = 10

# MODEL UTILITIES - FUNCTION 4 - STYLE, CONTENT, & TOTAL LOSS FUNCTIONS

# Now we define our loss functions
def style_content_loss(outputs):
  style_outputs = outputs['style']
  content_outputs = outputs['content']


  # Equation 5
  # Define the style loss function
  style_loss = tf.add_n([tf.reduce_mean(tf.square(style_outputs[name] - style_features[name])) for name in style_outputs.keys()])
  # Multiply the style loss by the weighted variable to get the weighted style loss
  style_loss *= beta / num_style_layers

  # Equation 1
  # Define the content loss function
  content_loss = tf.add_n([tf.reduce_mean(tf.square(content_outputs[name] - content_features[name])) for name in content_outputs.keys()])
  # Multiply the content loss by the weighted variable to get the weighted content loss
  content_loss *= alpha / num_content_layers

  # Equation 7
  # Define the total loss function
  total_loss = style_loss + content_loss
  return total_loss

# State the total variational weight
total_variational_weight= 30

@tf.function()
def train_step(img):
  # Tensorflow's GradientTape function performs automatic differentiation of the input
  with tf.GradientTape() as tape:
    outputs = feature_extractor(img)
    loss = style_content_loss(outputs)
    loss += total_variational_weight * tf.image.total_variation(img)

  # Apply the gradient by passing the loss and the generated image
  grad = tape.gradient(loss, img)

  # The gradient is optimized using the ADAM optimizer we declared earlier. This optimizes the
  # generated image using the gradient values
  optimizer.apply_gradients([(grad, img)])

  # The image is rewritten, being sure that the values all stay within the viable range of [0, 1]
  img.assign(clip_range(img))

import time
from IPython import display
# Start the run time
start = time.time()

epochs = 100
iterations = 100

step = 0
for n in range(epochs):
  for m in range(iterations):
    step += 1
    train_step(generated_img)
    print(".", end=' ')
  display.clear_output(wait=True)
  display.display(IU.deprocess_img(generated_img))
  print("Train step: {}".format(step))

end = time.time()
print("Total Time: {:.1f}".format(end-start))

file_name = 'output/stylized-image.png'
IU.deprocess_img(generated_img).save(file_name)
```

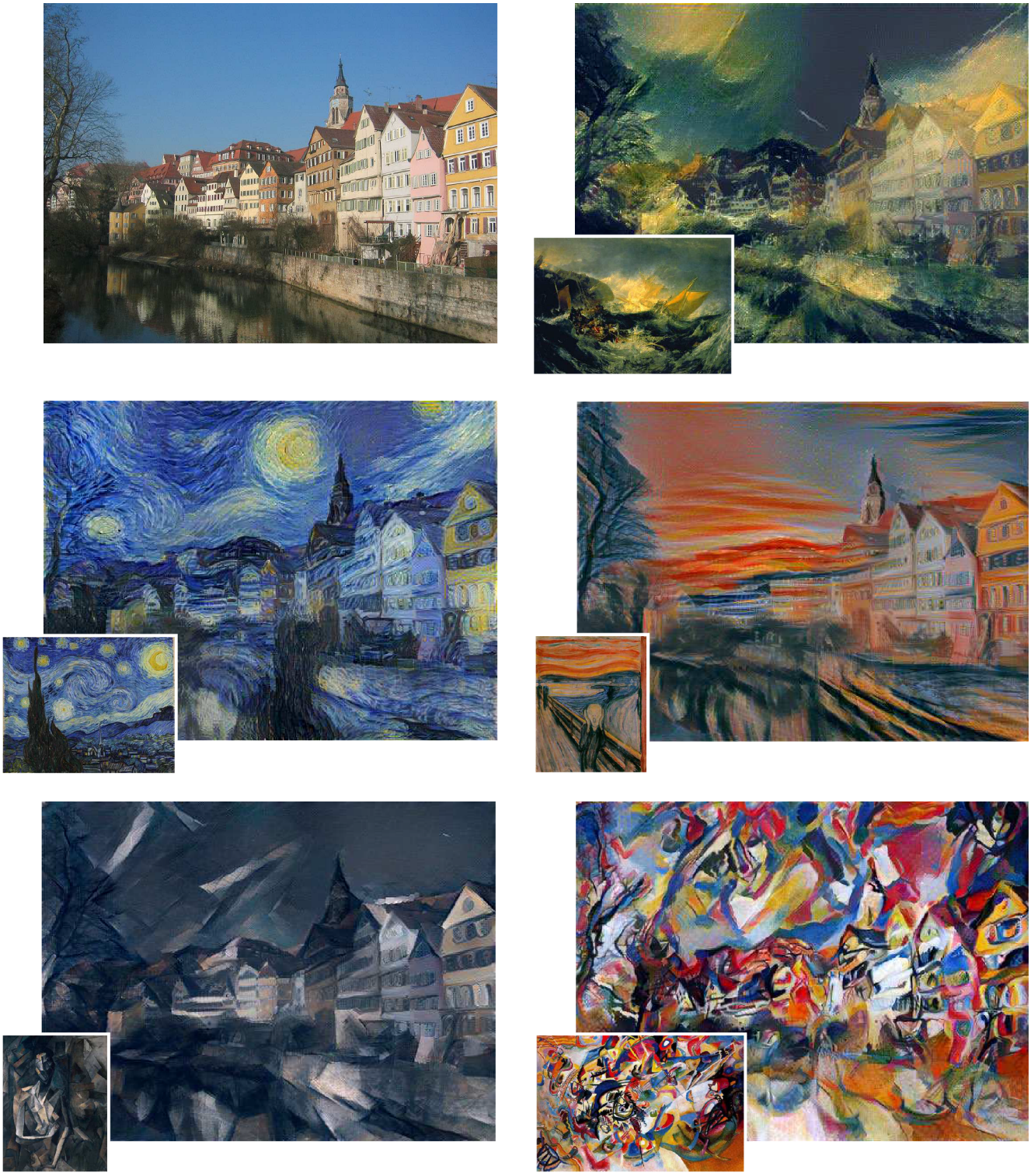

Figure 9. Theoretical Results from Gatys et al. The images were created by finding an image that simultaneously matches the content representation of the photograph and the style representation of the artwork. The original photograph depicting the Neckarfront in Tubingen, Germany, is shown (from the top left downwards then across to the right side) in **A** (Photo: Andreas Praefcke). The painting that provided the style for the respective generated image is shown in the bottom left corner of each panel. **B** The Shipwreck of the Minotaur by J.M.W. Turner, 1805. **C** The Starry Night by Vincent van Gogh, 1889. **D** Der Schrei by Edvard Munch, 1893. **E** Femme nue assise by Pablo Picasso, 1910. **F** Composition VII by Wassily Kandinsky, 1913.

Figure 10. Actual results from implementation. The original photograph depicting the skyline in Toronto, Canada, is shown (from the top left downwards then across to the right side) in **A** (Photo: Andreas Praefcke). The painting that provided the style for the respective generated image is shown in the bottom left corner of each panel. **B** The Shipwreck of the Minotaur by J.M.W. Turner, 1805. **C** The Starry Night by Vincent van Gogh, 1889. **D** Der Schrei by Edvard Munch, 1893. **E** Femme nue assise by Pablo Picasso, 1910. **F** Composition VII by Wassily Kandinsky, 1913.

VII. Results, Discussions, and Future Work

A. Theoretical Results

The results achieved by Gatys et al is shown in Figure 9. There are a couple of interesting differences to note. First and foremost, while it was very easy to derive the generic steps, there are some nuances that Gatys does not explicitly state in his paper. One of these is how many iterations are ran per result. The number of iterations dictates how much of the "style" is transferred over to the newly synthesized image. However, this can be remedied easily as it is very obvious to note after some iterations when the loss function converges. However, Gatys never stated what size his initial or final images were, despite his suggestion of constraining the image size before passing it through the VGG19 network.

The second is the optimizer. Gatys suggest using the L-BFGS optimizer to optimize the gradient of the total loss function with respect to the generated image. However, as discussed previously, the L-BFGS optimizer is not implemented in Tensorflow. Furthermore, the parameters of the optimizer were not stated.

The third nuance is that Gatys extracted the content image on layer 4_2. While it is never explicitly stated why this was done, it is assumed that Gatys assumed that extracting the content features from this layer was high enough to allow the more profound "features" of the content image to be displayed in the style transfer.

Finally, while the results from Gatys' style transfer serve as the benchmark, there is no numerical results to quantify how well his method works. As such, the results are evaluated just as art would be in real life: objectively.

B. Experimental Results

The implemented results from my code can be viewed in Figure 10. This section will discuss the deviations and remedies to the nuances discussed earlier. First and foremost, our images were synthesized with the following benchmark parameters

- Maximum Image Dimension: 1024 pixels
- Iterations/Epoch: 100
- Epochs: 100
- Optimizer: ADAM
- Learning Rate: 0.02
- Beta: 0.99
- Epsilon: 1e-1
- Alpha (Content Weight): 1e4
- Beta (Style Weight): 10
- Content Layers: 5_2

These parameters are important to note as some deviate from Gatys' original suggestions and others are derived from experimentation due to the lack of information provided in the paper. As stated previously, Gatys never specified the number of iterations or the image size. The maximum image size of 1024 was chosen as it is a staple size when it comes to images, as well as the fact that it is substantially large enough that the contents of the image are distinguishable when viewed. After some experimentation with the benchmark parameters, it was decided that 100 epochs with 100 iterations provided a result that was satisfactory.

The most notable deviation from Gatys' method was the difference in where the content image was extracted. Instead of extracting the content image on Gatys' suggested layer of 4_2, our content image was extracted on layer 5_2. This was due to the difference in the optimizer used in our algorithm. While it was possible to extract our content image on the same layer Gatys did, the results were not always pleasant or consistent. By that, it's meant that the model needs to be constantly adjusted based on the style image and content image used. It was fairly difficult to find the right balance of style/content weights and optimizer parameters that would yield suitable results from this layer. We also used the ADAM optimizer that was tuned with a learning rate of 0.02, a beta value of 0.99, and an epsilon of 1e-1. Using Gatys' Alpha and Beta values also yielded poor results, as we basically had to siwtch his values. In Gatys' method, the style weight was weighed significantly more that the content weight, ours was the complete opposite.These values were achieved after some investigation online. These were the hyper parameters that contributed to the deviation of our results from Gatys'.

The final deviation that our method has is that it accounts for total variational loss. While this was never mentioned in Gatys' paper, the total variational loss allows for the elimination of high frequency artifacts from our synthesized image, making them appear more natural.

Overall, our method seems to work quite well, although not as well as Gatys'. The majority of the images are comparable to the original results, however, the deviations become apparent when comparing images synthesized from Van Gogh's Starry Night. Gatys' results yielded a comparable image, with Van Gogh's swirled portrayal of the night sky as being more prominent, as well as the actual moon being included in the final image. Our result does a good job of copying this swirl pattern but not to the same degree. Furthermore, the patterns of the stars are not as intelligently placed as Gatys' results are. Surprisingly enough, Starry Night was used as the default image and is what was used to test the parameters found above. Other images were very easy to tune regardless of the parameters but Van Gogh's Starry Night had a very small window of passable results, which prompted the final benchmark values seen above.

C. Future Work

First and foremost, any future work done on this project should focus on trying to implement Gatys' method as closely as possible. While most modern and up to date tools make it easy to get close to Gatys' method, there are elements that are

lacking and crucial to obtaining the same results. Specifically, the optimizer used for gradient descent plays a larger role thatn originally understood. Despite my best efforts, implementing the L-BFGS optimizer was out of the scope for this project as it would require a much deeper understanding of numerical analysis that what this project entailed. Furthermore, many online resources have stated that using the ADAM optimizer was ”comparable” to using the L-BFGS optimizer, it was never truly elaborated on what that meant. While the results obtained can be called comparable to Gatys', there are still glaring issues that need to be addressed.

Furthermore, the hyper parameters used could also be fine tuned. With the use of a more powerful machine and parallel computing, different ranges and combinations of these parameters can be run in parallel to see what would give the closest result to the original.

There was also an interest in writing a script that produces a high resolution stylized image. The framework of this is to essentially start from a low resolution input, with a maximum dimension of 256 pixels, run that through the algorithm for a certain number of iterations, save the result and then use this result as the generated image for a secondary run with a maximum dimension of 512 pixels. After each run, the process would be repeated, with the result of the previous run being used as the generated initial image and continuously up-scaling the output until the desired high resolution result was achieved. The issue with this is that it would be fairly time consuming as the number of iterations would need to change every time, the hyper parameters tuned after every run, and the run times would increase exponentially as the resolution of the image increased. Current implementations exist and utilize 4 GPUs which result in an image with a maximum dimension of 3620 pixels.

The most impressive improvement to this algorithm would come in the form of either *photorealistic style transfer* or *multiple style transfer*. While the prior was not researche during this project, the latter already has an implementation. With a bit of tweaking, this code could accommodate multiple style images that would be used for the style transfer. The only constraint is optimization of

References

[1] L. Gatys, A.Ecker, M.Bethge, “Image Style Transfer Using Convolutional Neural Networks,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2414-2423.