Impact of Loss Functions and Model Feature Selections on CNN-Based Image Style Transfer Similarity Metrics

Richard Lepkowicz, Hannah Pavlovich
Georgia Institute of Technology
Atlanta, GA
rlepkowicz3@gatech.edu, hpavlovich3@gatech.edu
⋆for contribution information see table 1

Abstract

This study investigates the effectiveness of neural style transfer techniques, focusing on enhancing the preservation of artistic styles and content, particularly in architectural images. We begin by replicating Gatys et al.'s method using the VGG-19 model and then explore various extensions to improve style transfer from architectural styles to cityscapes and portraits. Our modifications include testing alternative loss functions—such as perceptual, total variation, and Wasserstein losses—to capture nuanced differences in style and content preservation. We also adjust CNN layer selections to assess how varying levels of feature abstraction affect transfer results to evaluate their impact on image quality.

Our results reveal that while the VGG-19 model provides a balanced integration of texture and content, the Perceptual Loss Method, though numerically promising, often produces simplistic overlays rather than effective style integration, especially with architectural images. The study concludes that high-level feature models generally offer a good balance but may struggle with complex imagery. Current evaluation metrics, such as cosine similarity, improve style-to-content ratios but may not fully capture qualitative aspects of style transfer. Further refinement in methods and metrics is needed to enhance style transfer for complex and varied imagery.

1. Introduction

In this study, we aim to explore and improve the process of style transfer, where we apply the artistic style of one image to another. Our primary goal is to understand how different models and similarity methods impact the effectiveness of style transfer. We specifically investigate how various techniques perform when applying highly textured architectural styles to different types of images, both figural and non-figural.

To achieve this, we experiment with a range of models and loss functions to determine which methods are most effective for transferring styles that are rich in texture compared to those that are more focused on content. We also examine the effects of using deep edges versus smooth edges and how the structure of the model and choice of loss functions influence the results. Our evaluation involves comparing how well each technique preserves the artistic style and maintains the original content. Ultimately, we seek to understand whether these factors matter and how they influence the quality of style transfer.

1.1. History and Current Practices

Style transfer has been around for several decades (early 1980s). The method has evolved from starting with digital signal processing to impart subtle artistic effects to photographs [1], such as blurring or softening, and progressed to the transfer of true artistic styles [2]. In the latter case, before the application of deep learning, a matched pair of a photograph and a painting of the same scene was needed to create a transfer map that could be applied to yet a third (new) image. The innovative work of Gatys et. al. [3] demonstrated the ability to transfer a wide range of artistic styles from paintings to photographs, particularly landscapes.

An inherent challenge of neural style transfer is how to quantitatively evaluate its effectiveness. Since the primary goal of neural style transfer is to generate visually appealing images by blending the content of one image with the style of another, the assessment is largely subjective. Traditional evaluation methods, which Gatys et. al. used, rely on human judgment for determining "goodness." This subjectivity presents a significant challenge in establishing a consistent and objective metric that can be used to test the effectiveness of different models and model parameters.

Student Name	Contributed Aspects	Details
Richard Lepkowicz	Model Building, Similarity Methods, Research	Built the base Style Transfer model and similarity score models. Implemented tests and analyzed results on different similarity methods. Researched loss methods.
Hannah Pavlovich	Modeling Building, Feature Selection Methods, Research	Developed similarity plot model. Implemented tests and analyzed results on different feature selections. Led research and paper formation.

Table 1: Contributions of team members.

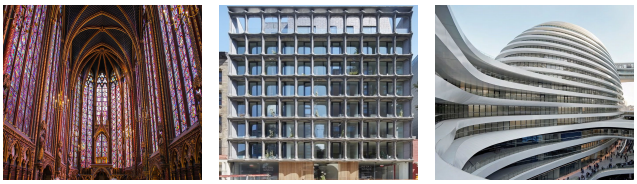
1.2. Study Implications

Improving style transfer techniques, particularly through the evaluation of different similarity metrics, holds significant potential for advancing various fields. This study focuses on the nuanced application of style transfer to architectural images and figural content, providing a novel perspective on how these techniques can be optimized. For architects and designers, enhanced style transfer methods can lead to more realistic and artistically textured renderings of buildings, transforming how architectural projects are visualized. This capability allows for photorealistic presentations that also incorporate rich, artistic textures, offering a new dimension to architectural design and visualization. Similarly, artists working with figural images can benefit from these advancements by integrating complex textures and styles, thereby expanding their creative possibilities and enhancing the visual appeal of their work.

Additionally, our work provides valuable insights for data scientists and practitioners by examining how different methods influence the balance between preserving content and applying style in the final image. Understanding the impact of various models and loss functions enables professionals to make informed decisions about which approaches best suit their specific goals, whether they aim to emphasize artistic style or maintain content fidelity. This guidance is crucial for applications across digital art, media, and design, where the ability to fine-tune the presence of style versus content can significantly impact the effectiveness and quality of the output. By refining these techniques and offering clear criteria for selecting models and metrics, we contribute to more versatile and effective tools for a wide range of real-world applications.

1.3. Data Selection

Gatys et al.'s method represents a major advancement from earlier non-parametric techniques, which transferred textures without considering feature hierarchies [3]. Using a convolutional neural network (CNN) like VGG-19, Gatys's approach extracts and combines style and content features from separate images. Their findings show that using more convolutional layers aligns the resulting image more closely with the style, while reducing the influence of the content structure. This demonstrates how low-level features, such as color edges, evolve into more detailed content representations.



Cathedral Xoco Zaha

Figure 1: Style Images

Miley Cyrus Cityscape

Figure 2: Content Images

Gatys's visuals reveal that lower layers retain distinct color boundaries, whereas higher layers capture more intricate brush strokes and style details. Geirhos et al. note that CNNs often over-classify textures at the expense of shapes, which can be mitigated through data augmentation [4]. Yosinski et al. further illustrate that early layers capture basic edges, while deeper layers reveal complex content [5]. These insights guide our selection of convolutional layers for style transfer experiments.

Our style images, figure 1, are chosen to test various levels of style transfer. "Cathedral," with its soft-edged stained glass, is ideal for examining low-level features. "Xoco," featuring a grid-like façade, provides a structured but less pronounced style for testing intermediate layers. "Zaha," depicting Zaha Hadid's Galaxy SOHO, presents strong contrasts and dynamic curves, offering a challenge for higher-level style integration.

For content images, figure 2, we use a headshot of Miley Cyrus and a cityscape. The headshot, being more figural, is suited for capturing detailed content, while the cityscape's high-level structures represent a more abstract content base. This selection aims to test how well style transfer handles different content types. Xing's research on portrait-aware style transfer, which uses masking to apply textures at different levels [6], offers additional context. However, our experiment will test various layer combinations without masking to identify the best balance between style and content.

2. Approach

For more information, see our GitHub repository: https://github.com/rlepko/final_project

This project aims to replicate the initial work of Gatys and then look at extensions/modifications of the base model to allow the style transfer of architectural images of buildings to portraits. Specifically, we will explore the following modifications to the original Gatys method:

1. *Different Loss Functions:* While the baseline model employs the RMS pixel loss, we will experiment with alternative loss functions such as perceptual loss, total variation loss, and Wasserstein loss. These loss functions are selected to potentially improve the quality of style transfer by capturing more nuanced differences in style and content preservation.

2. *CNN Layer Selection:* The choice of CNN layers for extracting content and style features significantly impacts the results of style transfer. We will modify the selection of layers used in the VGG-19 network to determine how deeper or shallower layers affect the transfer results. This helps in understanding the role of different levels of feature abstraction in style transfer.

2.1. Selection of Evaluation Metric

To address the need for a more objective evaluation metric we performed a literature search to see what has been previously tried to make the analysis of neural style transfer more objective. We explored both the Structural Similarity Index (SSIM) [7] and the Inception Score [8] methods. After reviewing the inception score method in more detail, we determined the method irrelevant since it is designed, in part, to determine the diversity of generated images, while our goal is to disambiguate between the fraction of the content and style component of the images.

The SSIM metric appeared attractive because it evaluates images in terms of the luminance, contrast, and structural content of the image, which are relevant parameters to a human observer. However, during our experiments, we found that SSIM did not adequately separate our architectural images from the celebrity images into distinct classes. While SSIM values were relatively high for celebrity images, likely due to their uniform backgrounds, the architectural images exhibited low SSIM scores. This discrepancy suggested that SSIM might not be a suitable metric for evaluating the success of style transfer, particularly when dealing with diverse and complex image structures.

To find a better method to separate the image classes, we turned to a feature-based approach using a pre-trained deep learning model. While we could not find a reference to this approach, an image classification network (that is separate and independent from the network used for the style transfer) has proven successful in transfer learning applications. We implemented the VGG-19 model, which has robust feature extraction capabilities, and by removing the classification layers, we extracted high-level features from the last convolutional layers of the model. These feature vectors represent key characteristics of the images including textures, edges, and shapes. We then applied Cosine similarity [9], which measures the cosine of the angle between two non-zero vectors allowing us to capture the high-level similarities and differences between images. Figure 3 shows an example of the types of images (architectural and portraits) used in the remainder of this analysis and a set of 6 of each Cosine similarity scores.

2.2. Implementation of Baseline Model

To develop the baseline model for our neural style transfer project, we followed the original method proposed by Gatys et al. in their seminal paper "A Neural Algorithm of Artistic Style." To gain a deeper understanding and practical implementation guidance, we referred to a YouTube tutorial [10], which provided a step-by-step explanation of how to implement the algorithm in Python using PyTorch. The primary difference noted in the YouTube tutorial was the omission of normalization layers, which, interestingly, did not significantly affect the results.

To verify our baseline model, we downloaded the same images used in the Gatys paper: a photograph of the Neckarfront in Tübingen as the content image and Vincent van Gogh's "The Starry Night" as the style image. Figure 14a (in the appendix) shows that we were able to successfully recreate the image like what was shown in the paper. Additionally, we saved the generated image every 200 iterations (we iterated over 6000 total steps) to demonstrate the utility of the evaluation metric. As can be seen if Fig. 14ac we observe the expect result that the generated image is much more like the content image and transitions over to the style image, but as intended it does not become a replica of the style image and levels off to a style value of approximately 0.4.

2.3. Implementation of CNN Architecture

We used three architectures to evaluate the models:

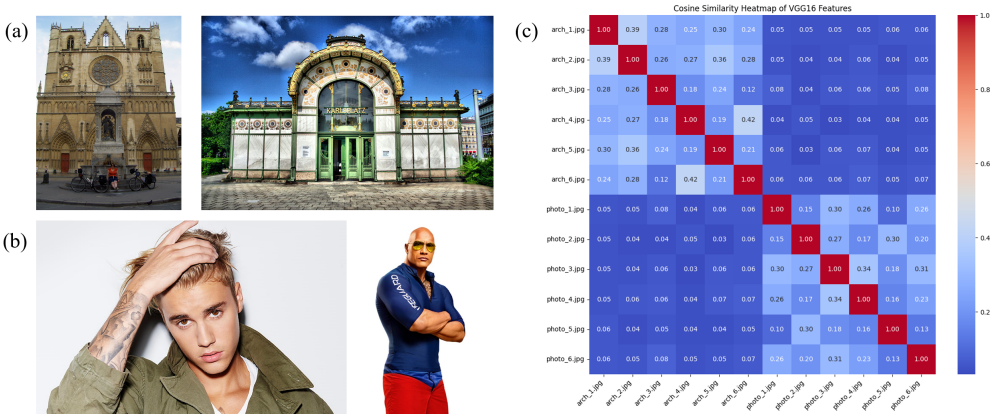


Figure 3: Sample of the types of (A) architectural and (B) portrait images used in this study as well as the (C) cosine similarity scores between the two classes of images

1. Standard: This architecture comes from the original Gatys text, and uses the layers in the original VGG-19 model shown in figure 9a in the appendix. Full connected layers are not included in the network, as the flattening of the input would lose the spatial representation of the layers.

2. Low Features: Selecting conv_1, conv_2, and conv_3 as seen in figure 12b extracts the low level features of the style image. We hypothesize that these low level features will work particularly well for zaha and xoco, which have strong edges.

3. High Features: Selecting conv_4 and conv_5 as seen in figure 12c selects the high-level features of the model, preferring content over style. We expect that this structure would favor cathedral, which has strong context without texture.

2.4. Implementation of Alternate Loss Methods

We explored two primary loss methods: the Perceptual model and the Wasserstein loss. The Perceptual model leverages a pre-trained VGG-19 neural networks to measure the similarity between the generated image and the content and style images based on high-level feature representations. The Wasserstein loss, on the other hand, measures the distributional distance between the generated image and the style image, promoting more realistic texture synthesis. For the sake of brevity, we will discuss the results of the Perceptual model in detail.

2.5. Expectations and Challenges

One key achievement of this project was successfully implementing a working model. While style transfer uses a straightforward CNN, implementation posed challenges. A helpful video tutorial enabled us to understand the methodology and develop the primary algorithm. Surprisingly, our models effectively transferred style from photographs, a feat many papers suggested was difficult, especially with architectural images.

Implementing our extensions was also challenging. Though not entirely novel, figuring out the concepts and tuning parameters was necessary. Experimenting with different layer structures worked well and required minimal hyperparameter tuning. However, despite correctly implementing the perceptual loss model, it failed to transfer style and only produced image overlays, highlighting the mixed results of our innovations.

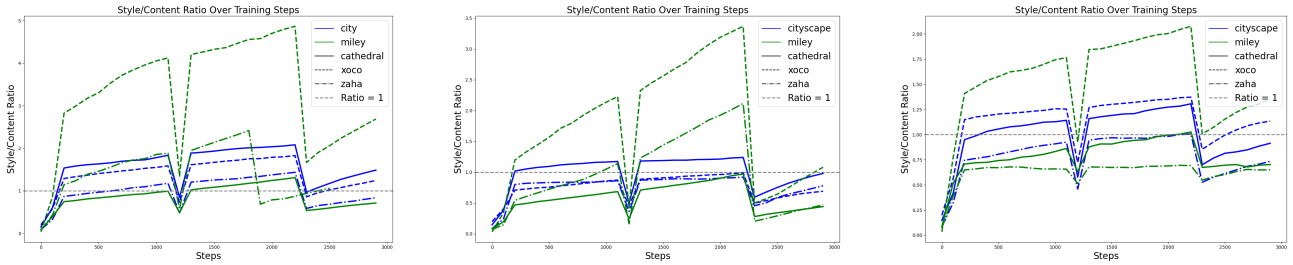
3. Experiments and Results

3.1. CNN Architecture Selection

Six images were created from each of the architectures described in section 2.3, with a selected few seen in figure 5. All resultant images along with larger versions of the plots can be seen in the appendix, figure 10.

The VGG-19 model, as described in the original Gatys paper, effectively balances texture and content in style transfer. The cathedral style appears nuanced and translates to what resembles colorful white noise when applied to the Miley Cyrus image, evident in the plot in 9a, which shows the generated image has a low style/content ratio. In contrast, the cathedral style applied to a cityscape maintains a more coherent representation of stained glass, which may be attributed to both images sharing a one-point perspective.

When applying the Xoco and Zaha styles, we observe different outcomes. The Xoco style retains the figure reading of both Miley Cyrus and the cityscape most effectively, while the Zaha style emphasizes the swooping forms of the



(a) standard VGG-19 (b) low-level features (c) high-level features

Figure 4: Cosine Similarity Plots

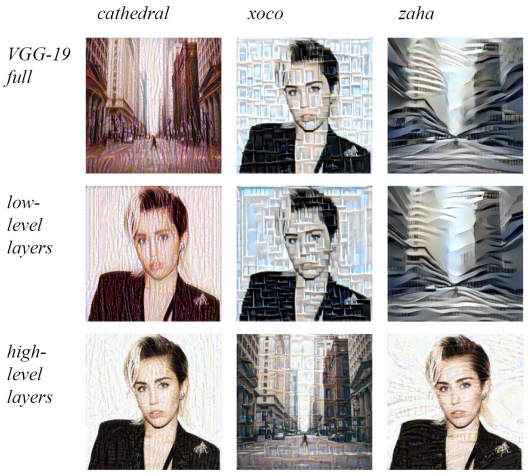


Figure 5: generated images after 3000 steps from different style/content pairings

content images. The Xoco style highlights strong edges, particularly visible in the cityscape, but tends to emphasize texture over color. In contrast, the Cathedral style, when compared to the full VGG-19 model, produces visually similar results, likely because the Cathedral style image lacks strong edges.

High-level features of the network, which ignore the low-level details and edges, create visually distinct results from those produced by the lower layers. These high-level features focus more on texture rather than edges, with the Cathedral style showcasing the most intricate details. The Miley Cyrus image remains relatively unchanged by these high-level features, while the cityscape undergoes more noticeable alterations. This difference could stem from the cityscape's inherent complexity, including various levels of content, edges, and motifs, whereas the Miley Cyrus image is more straightforwardly figural.

The similarity plots provided in the appendix (figure 11) and the ratio of the style to content similarity scores are shown in figure 4 offer valuable insights. For Miley Cyrus with the Xoco style, all models exhibit a significant bias towards style over content. However, with high-level features, the style-to-content ratio decreases, suggesting that the program struggles to capture texture without the lower layers.

The Cathedral images consistently show a higher content-to-style ratio. In the low-level features, the style-to-content ratio never exceeds one, supporting the notion that the Cathedral style is not effectively conveyed through low-level features alone.

Finally, the cityscape images demonstrate a more consistent style-to-content ratio compared to Miley Cyrus images, regardless of texture. This consistency may be due to the cityscape's inherent complexity or the greater similarity between the cityscape and architectural style images. In contrast, the initial dissimilarity of the Miley Cyrus image leads to greater polarization in the style-to-content ratio.

3.2. Perceptual Loss Method

We implemented the Perceptual model to enhance neural style transfer by leveraging high-level feature representations from a pre-trained neural network. Specifically, we used the VGG-19 model, a well-established convolutional neural network trained on the ImageNet dataset, to extract features from the content and style images. The VGG-19 model's convolutional layers are adept at capturing various levels of abstraction in an image.

The Perceptual model differs fundamentally from the base Gatys loss method, which relies on pixel-wise comparisons. Instead of comparing images at the pixel level, the Perceptual model computes the loss based on the similarity of high-level features extracted by the VGG-19 model. This is done by passing the content, style, and generated images through the VGG-19 network and extracting features from specific layers known to capture different levels of image abstraction. These layers include early layers that capture edges and textures and deeper layers that capture more complex patterns and structures.

The content loss is calculated as the mean squared error between the features of the generated image and the content image, while the style loss is computed as the mean squared error between the Gram matrices of the generated image and

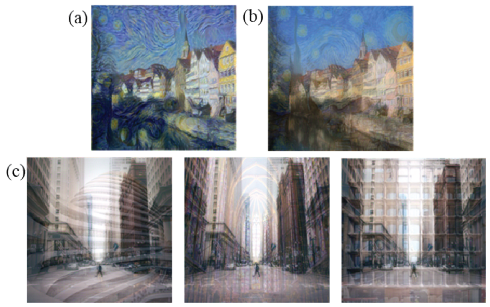


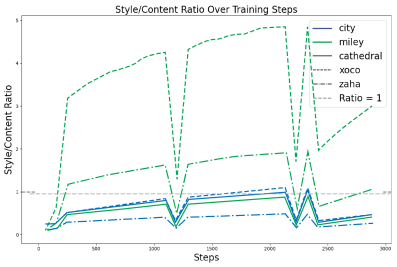
Figure 6: (a) Style transfer for the loss method in the original Gatys paper compared to the (b) style transfer for the developed perceptual model contrasted with the (c) same model applied to our selected cityscape image. See appendix for additional images (figure 13)

the style image. The Gram matrix, which captures the correlations between different feature maps, is a crucial component in capturing the style of an image. By using these high-level features and their correlations, the Perceptual model aims to produce more visually coherent and stylistically accurate results compared to pixel-wise loss methods.

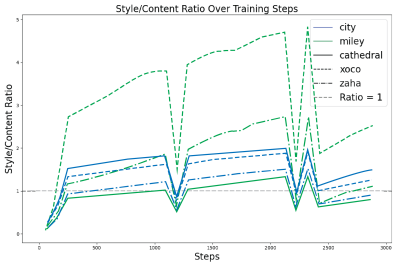
To achieve effective style transfer using the Perceptual model, we found that a much larger learning rate was required. Additionally, the weights for the content and style losses, denoted as alpha and beta, needed to be balanced differently than in the base Gatys method. Specifically, setting both alpha and beta to equal values yielded the best results. This adjustment underscores the difference in how the Perceptual model evaluates the importance of content and style features compared to the base method.

When tested with the base Gatys images, such as transferring the style of Van Gogh's "Starry Night" onto a village scene, the Perceptual model performed reasonably well (figure 6a,b). The resulting images showed a good balance of content preservation and stylistic transformation, demonstrating the model's capability. However, the Perceptual model struggled with our architectural image set (figure 6c). Instead of seamlessly blending the content and style, the generated images appeared as simple overlays of the two inputs, failing to achieve the desired stylistic integration.

Despite the promising results with the Gatys images, the Perceptual model did not generalize well to our architectural images. This discrepancy highlights the importance of selecting appropriate style and content pairs and suggests that further refinement and perhaps domain-specific adjustments are necessary to improve performance with more complex and varied image sets. Figure 14b shows the Cosine Similarity metric for both the base model and the perceptual model which approximately supports this result showing a more significant transfer of style (larger ratio) for the base model compared to that of the perceptual model.



(a) base loss model



(b) perceptual loss model

Figure 7: The ratio of the style to content similarity metric developed in Section 2.2

4. Conclusion

Our study explores the effectiveness of CNN architectures and loss methods in style transfer. High-level feature models generally provide the best balance between style and content, focusing on abstract features for better integration. However, visually, these models often fall short, particularly with complex or photorealistic images, where the results lack stylistic coherence.

The Perceptual model, which uses high-level features, shows improved numerical style-to-content ratios but struggles with architectural images, often resulting in simple overlays rather than effective style integration. This suggests that the Perceptual model may not be ideal for complex or photorealistic scenarios.

Overall, while numerical metrics indicate that high-level feature models and perceptual loss methods offer balanced style-to-content ratios, these scores do not always align with visual quality. The better numerical ratios often correspond to images that resemble overlapping rather than true style transfer. Specifically, the Cathedral style, despite having the lowest style-to-content ratio and blending in visually, highlights that the metrics may favor less pronounced styles. This suggests that both style and content might be obscured rather than effectively represented. Thus, our findings emphasize the need for a comprehensive approach that considers both numerical and visual evaluations to fully capture the effectiveness of style transfer methods.

References

[1] Aaron Hertzmann. “Image Stylization: History and Future (Part 1)”. In: *Adobe Research* (). URL: https://research.adobe.com/news/image-stylization-history-and-future/.

[2] Aaron Hertzmann et al. “Image Analogies”. In: *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2023. ISBN: 9798400708978. URL: https://doi.org/10.1145/3596711.3596770.

[3] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. “Image Style Transfer Using Convolutional Neural Networks”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 2414–2423. DOI: 10.1109/CVPR.2016.265.

[4] Robert Geirhos et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness”. In: *CoRR* abs/1811.12231 (2018). arXiv: 1811.12231. URL: http://arxiv.org/abs/1811.12231.

[5] Jason Yosinski et al. “Understanding Neural Networks Through Deep Visualization”. In: *CoRR* abs/1506.06579 (2015). arXiv: 1506.06579. URL: http://arxiv.org/abs/1506.06579.

[6] Yeli Xing et al. “Portrait-Aware Artistic Style Transfer”. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. 2018, pp. 2117–2121. DOI: 10.1109/ICIP.2018.8451054.

[7] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE Transactions on Image Processing* 13.4 (2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[8] Tim Salimans et al. “Improved Techniques for Training GANs”. In: *CoRR* abs/1606.03498 (2016). arXiv: 1606.03498. URL: http://arxiv.org/abs/1606.03498.

[9] Amit Singhal and I. Google. “Modern Information Retrieval: A Brief Overview”. In: *IEEE Data Engineering Bulletin* 24 (Jan. 2001).

[10] Aladdin Persson. *Pytorch Neural Style Transfer Tutorial*. 2020. URL: www.youtube.com/watch?v=imX4kSKDY7s&t=632s.

Appendix

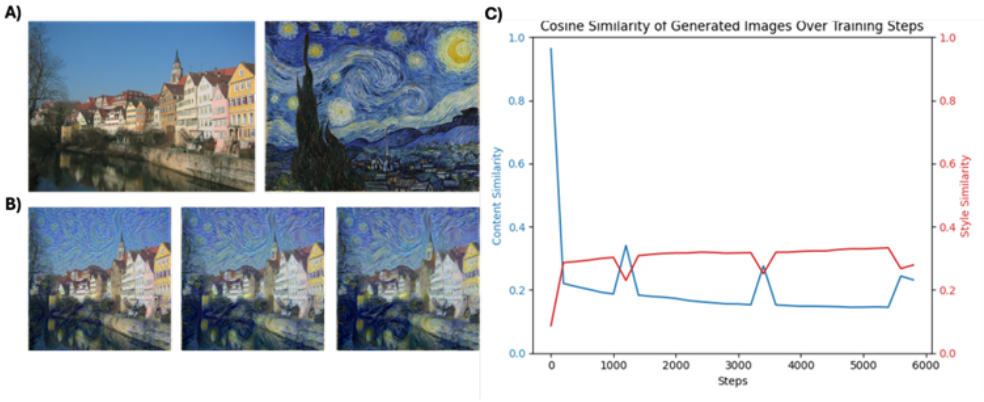
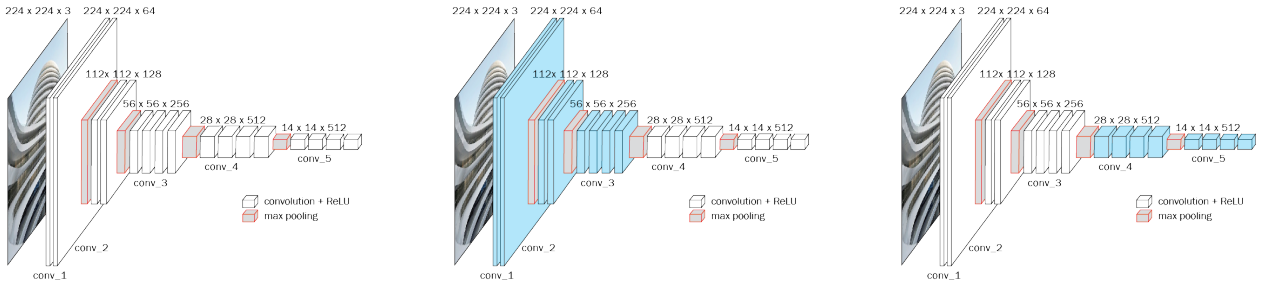


Figure 8: Replication of baseline model presented in Gatys [3] using the (A) sample images from the paper and (B) the generated images at 1000, 3000, and 6000 iterations (C) the cosine similarity scores for the generated image to both the content and the style image over the 6000 iterations.Two figures side by side



(a) standard VGG-19 (b) low-level features (c) high-level features

Figure 9: VGG-19 Architecture with selected layers in blue

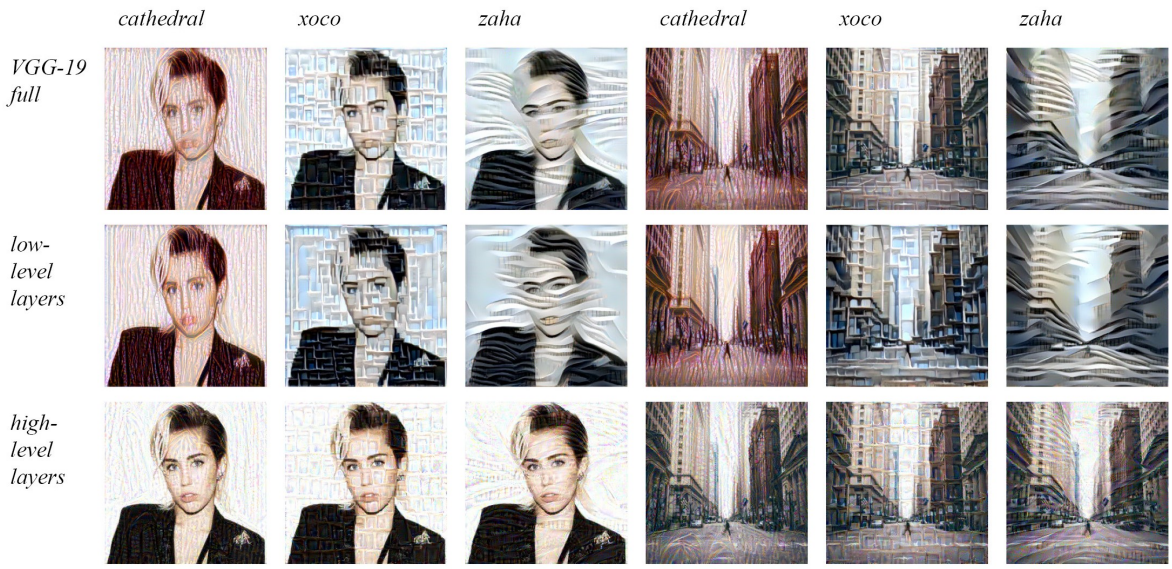


Figure 10: generated images after 3000 steps from different style/content pairings and feature selections

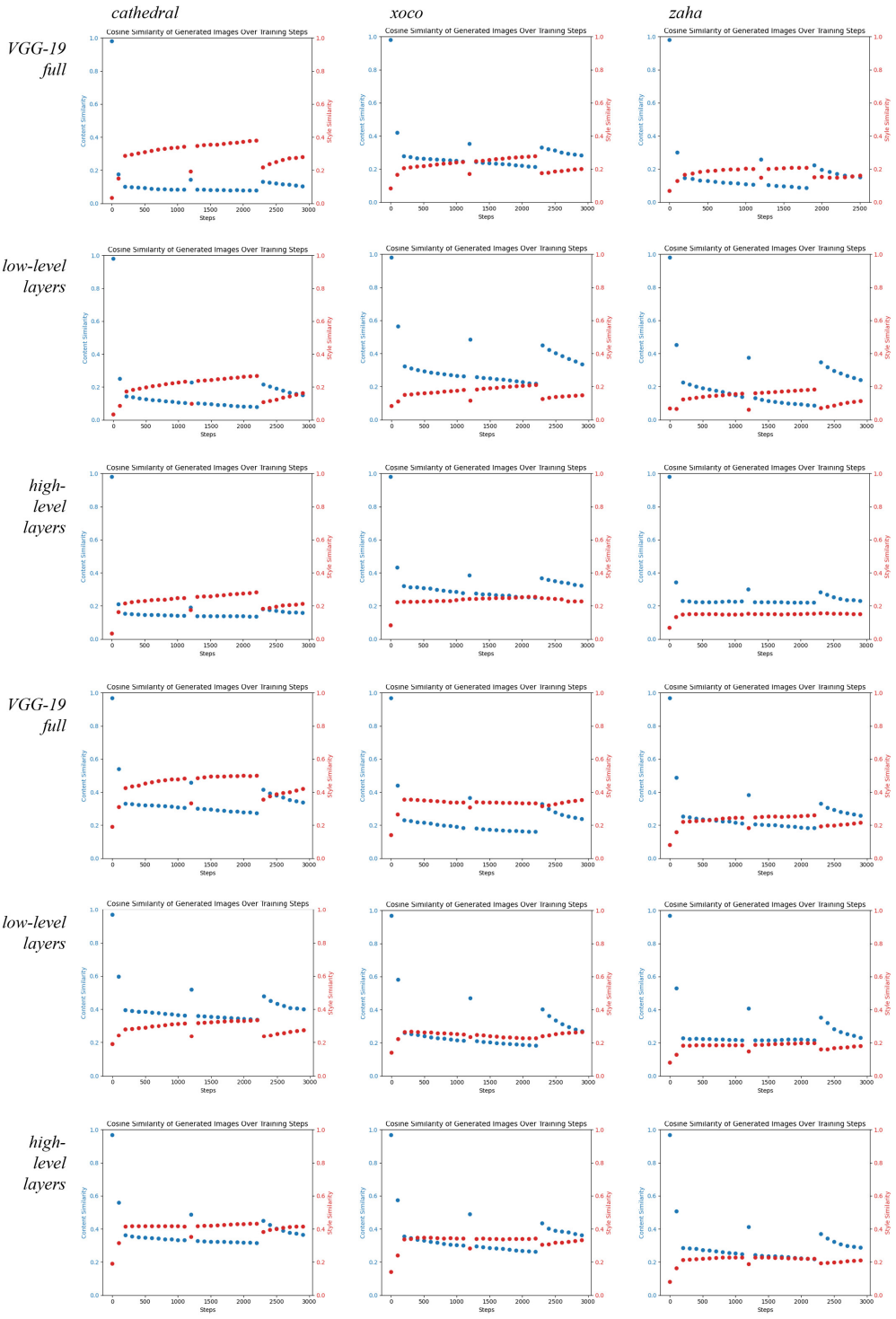
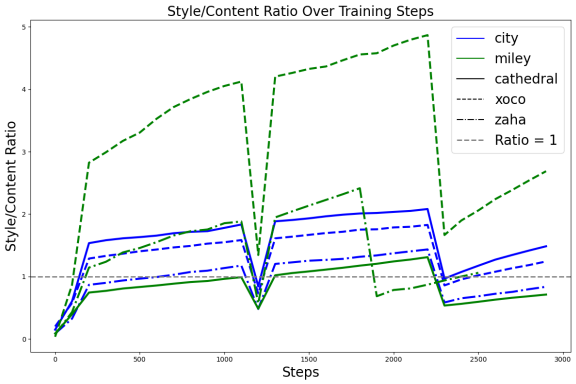
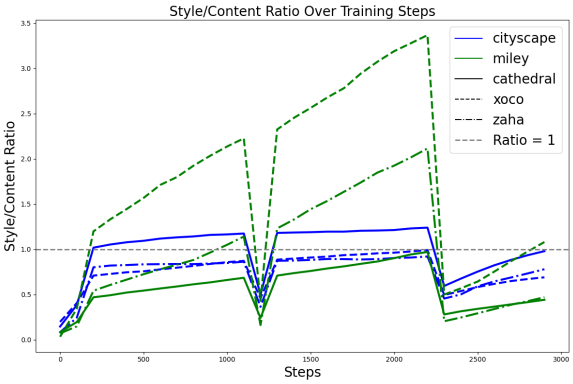


Figure 11: Plots of similarity from Style Transfer generated images to 3000 steps. Blue represents the content similarity to the generated image, and red represents the style similarity to the generated image.



(a) standard VGG-19



(b) low-level features

(c) high-level features

Figure 12: The ratio of the style to content similarity metric for feature selection as developed in Section 3.1

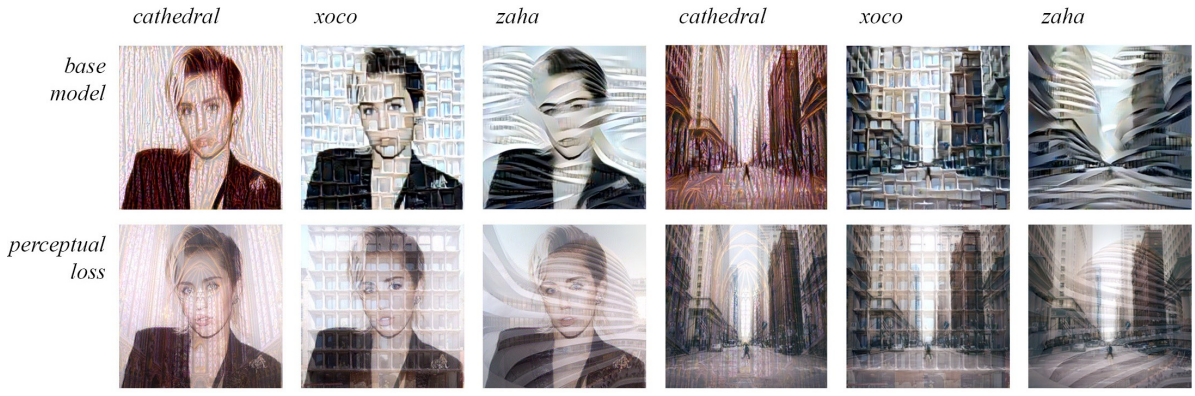
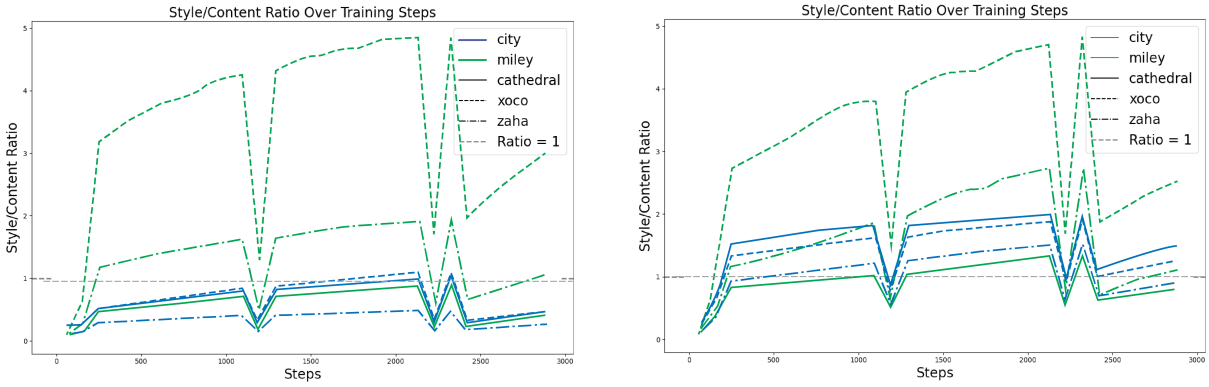


Figure 13: generated images after 3000 steps from different style/content pairings and loss model function



(a) base loss model

(b) perceptual loss model

Figure 14: The ratio of the style to content similarity metric for loss functions as developed in Section 2.2