Exploring Neural Style Transfer with Alternative CNN Architectures

CaiYi Pei, Yi Dai, Zheng Chen

April 19, 2024

1 Introduction

Neural Style Transfer blends art and technology by leveraging machine learning to merge the stylistic features of one image with the content of another. Traditionally, this process has leaned on the VGG network for its robust feature extraction. However, this project delves into alternative neural network architectures such as MobileNet, EfficientNet, and Inception-v3 to enhance style transfer. These networks offer potential new insights and improvements by learning and replicating complex patterns, crucial for creating visually engaging artwork. This exploration not only enriches artistic workflows but also expands the expressive capabilities of artists, empowering them to innovate and redefine the art creation process(e.g An artisitc cat can be generated from pictures of a cat and a tree in Appendix Figure 12). Investigating these alternative architectures represents a promising direction with significant potential to impact artistic expression.

2 Background and Related Work

The foundational work for this project is Gatys et al.'s "A Neural Algorithm of Artistic Style", which pioneered the use of CNNs to replicate the style of famous paintings onto photographs using a pretrained VGG-19 network[1]. This seminal paper established the groundwork for using deep neural networks for artistic style transformation by utilizing a feature space designed to capture textural information. Since this pioneering work, subsequent studies have refined the approach, yet have largely remained confined to similar architectures. The exploration of alternative architectures in neural style transfer remains underexplored, offering a rich vein of potential advancements in the field.

In addition to VGG-19, other CNN architectures have shown promise in various applications:

- **EfficientNet**[2]: EfficientNet introduces a novel approach called "compound scaling," which optimally scales width, depth, and resolution of the network to achieve remarkable efficiency without compromising accuracy. This adaptability makes EfficientNet suitable for various computational budgets and hardware capabilities.

- **MobileNet**[3]: MobileNet is a streamlined convolutional neural network tailored for mobile vision applications. Its low computational intensity makes it ideal for real-world applications like object detection, fine-grained classifications, and localization. MobileNet utilizes depthwise convolutions to drastically reduce parameters, resulting in a lightweight deep neural network.

- **Inception-v3**[5]: Inception-v3, developed by Google, incorporates enhancements such as factorized convolutions, auxiliary classifiers, and extensive use of batch normalization to balance efficiency and accuracy. It excels in image recognition tasks without excessive computational demands and is widely adopted across various computer vision applications, including object detection, segmentation, and style transfer, due to its effective handling of multi-scale information.

3 Data Processing

For neural style transfer, while traditional image datasets are not directly employed, the use of pretrained models requires a foundational dataset. In this project, the CNN architectures are pretrained on ImageNet and are minimally fine-tuned, especially in their later layers, to effectively capture the artistic nuances essential for high-quality style transfer. This ensures the models maintain their proficiency in recognizing complex visual patterns while adapting to specific artistic styles.

A substantial portion of our images comes from the National Gallery of Art, which offers its collection digitally under the Creative Commons Zero (CC0) license, allowing free use without copyright concerns.

Input images are converted to 'RGB' color space and resized to the model-specific dimensions (299x299 for Inception-v3 and 400x400 for other models). These images are then converted into PyTorch tensors and normalized using mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225] to match the preprocessing standards of ImageNet training. To check the effectiveness of preprocessing visually, we use an imshow function that reverses the normalization, allowing for precise assessment and refinement of our preprocessing techniques.

4 Architecture

Our project leverages Convolutional Neural Network (CNN) architectures, specifically Inception-v3, EfficientNet, and VGG19, as feature extractors for style transfer. These models are renowned for their unique capabilities in handling intricate patterns and structures within images. Inception-v3 is utilized for its multi-scale processing through inception modules, which is expected to maintain content detail while integrating the desired artistic style. EfficientNet's systematic scaling enables capturing a diverse set of style features, potentially enriching the transferred style's complexity and depth.

We harness selected layers within these architectures to perform optimization on the transferred image. The optimization objective is defined as the mean squared error (MSE) between the feature representations of the selected layers of the target image and those of the style and content images, differentiating between content loss and style loss.

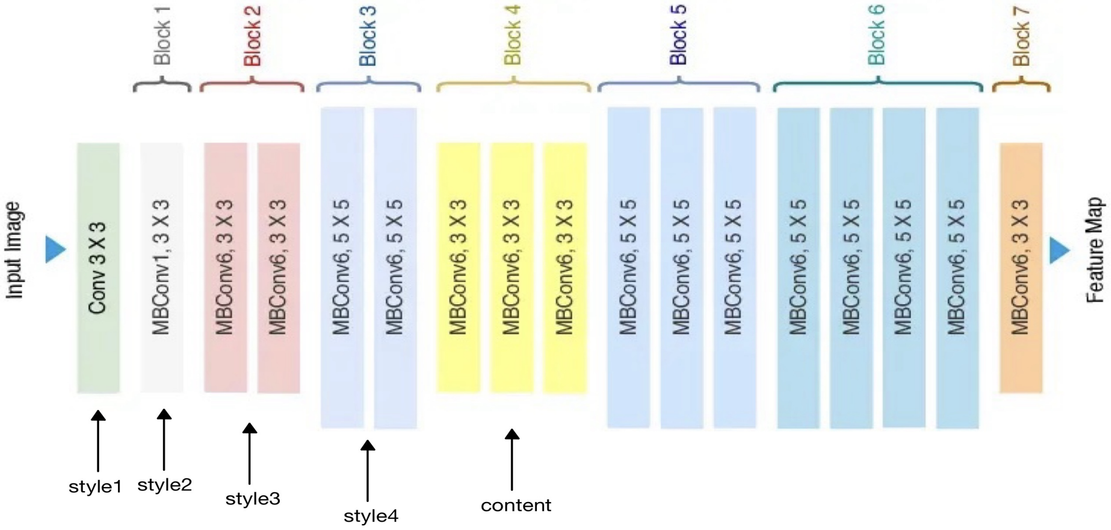


Figure 1: Selected Layers from EfficientNet for NST [6]

Due to constraints in computational resources and time, the training process begins from a copy of the content image rather than a noise image, which is a common practice when resources are limited. This approach often accelerates convergence and requires less computational effort, albeit at the potential cost of introducing less variation in the initial stages of the style transfer process. Our methodology is designed to strike a balance, optimizing the quality of the artistic output while working within the confines of available computational capacity.

5 Baseline Model

The baseline for this project is the traditional VGG19-based style transfer model as originally introduced by Gatys et al. in their seminal work. VGG19, known for its effectiveness in deep feature extraction due to its deep convolutional layers, utilizes the Gram matrix to capture and represent artistic style features. This model serves as a benchmark to compare the performance of alternative architectures explored in this project. Our goal is to assess whether newer models can surpass VGG19 in terms of processing speed, image fidelity, and overall aesthetic quality. We will compare our results to the second implementation of this original paper available on CatalyzeX, which adheres closely to the established VGG19 style transfer methodology [1][7].

6 Result

6.1 Quantitative Results

Content	Style	VGG-19	EfficientNet	Mobile	Inception-V3
Artemis	Glass 1	19417.9	1.86e13	2174.3	659.4
Artemis	Glass 2	72095.9	3.20e13	884.2	471.8
Artemis	Glass 3	47479.9	8.56e13	2337.1	8459.5
Artemis	Glass 4	15700.5	1.58e13	465.9	1368.4
Einstein	Artemis	67313.6	8.10e13	3643.7	6503.8
Einstein	Christ	138040.3	1.63e13	533.3	15608.4
Einstein	Yamada	477484.0	1.61e14	1777.0	23656.7
Einstein	Lucy	9765739.0	1.55e14	1037.9	14541.1

Table 1: Final Total Loss

6.2 Qualitative Results

In the qualitative analysis of the outputs generated by our selected neural networks, each architecture revealed distinct tendencies in style transfer, particularly when negotiating the complexity of background elements.

When presented with a content image set against a pure or minimalist background, all models exhibited a robust ability to perform style transfer with high aesthetic appeal(see Figure 1-4, higher resolution picture can be found in our github repo). Good performance in this context is characterized by the successful capture of the style image's texture and pattern, while maintaining the general shape and recognizable features of the content image. The pure background facilitates a clearer transfer of style, allowing the models to blend elements without the distraction of complex background patterns or colors, aligning with findings from studies such as those on the Artemis dataset.

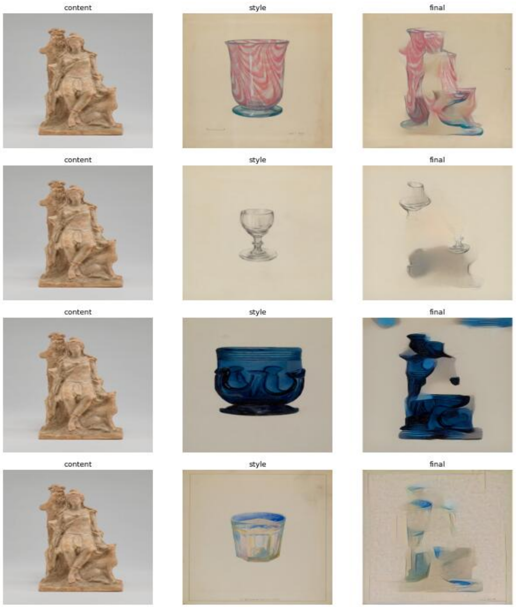


Figure 2: VGG-19, clean background

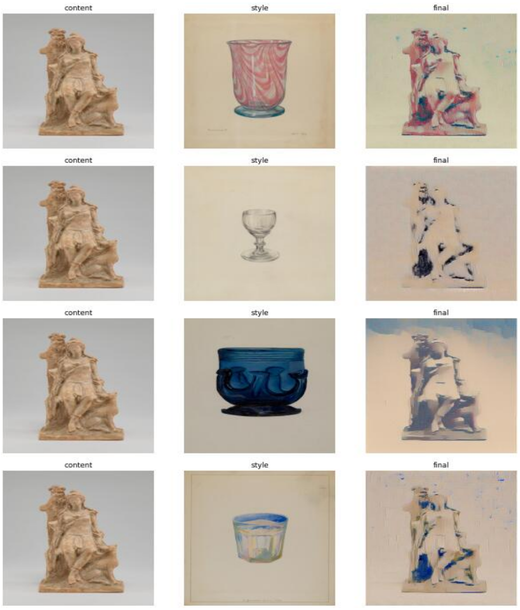


Figure 3: EfficientNet, clean background

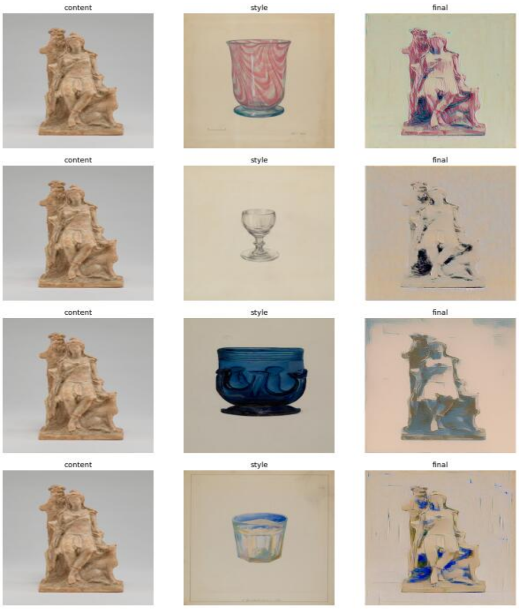


Figure 4: Mobile, clean background

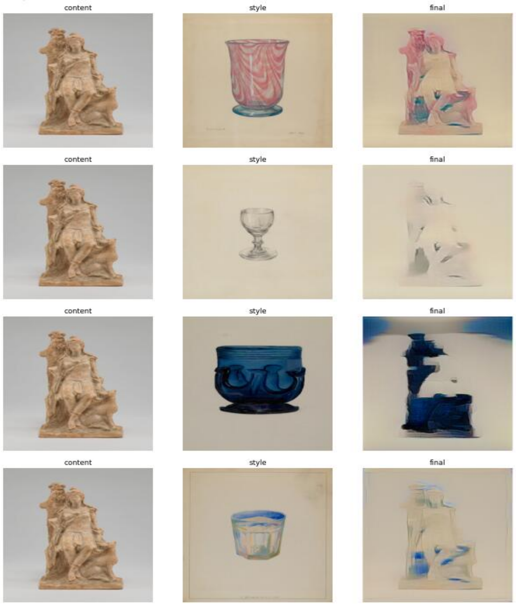


Figure 5: Inception-V3, clean background

Upon increasing the complexity of the background, however, model-specific behaviors emerged.

VGG19 frequently distort the content, particularly when backgrounds are complex. This could result in artistic outputs that, while potentially more abstract, might diverge from the original content's structure. VGG19 might be best employed when an artistic reinterpretation of the content is the goal, rather than a faithful preservation of content detail.

EfficientNet and MobileNet achieved a moderate and balanced style transfer. MobileNet adeptly navigated the trade-off between style and content, offering a compromise that often retained content integrity without under-representing the style image's characteristics. This balance makes MobileNet a suitable generalist option for style transfer tasks across a range of background complexities. On the other hand, the style transfers from EfficientNet occasionally suffered from poor integration of style elements, leading to less coherent outputs. Although it is designed to efficiently scale model size and manage computational resources, EfficientNet still exhibited suboptimal results in our tests.

Inception-v3, with its capacity for detailed multi-scale information processing, tended to over-apply stylistic features. This sometimes resulted in the overshadowing of content details, which may not be desired in cases where content preservation is paramount. Nevertheless, for applications where a bold stylistic expression is favored, this tendency could be harnessed creatively.

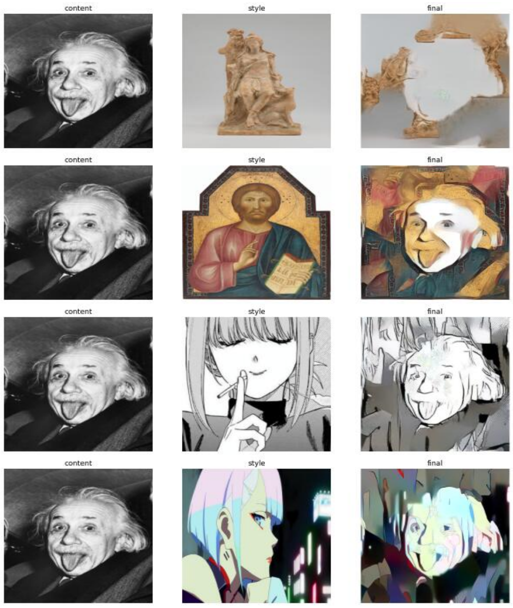


Figure 6: VGG-19, complex background

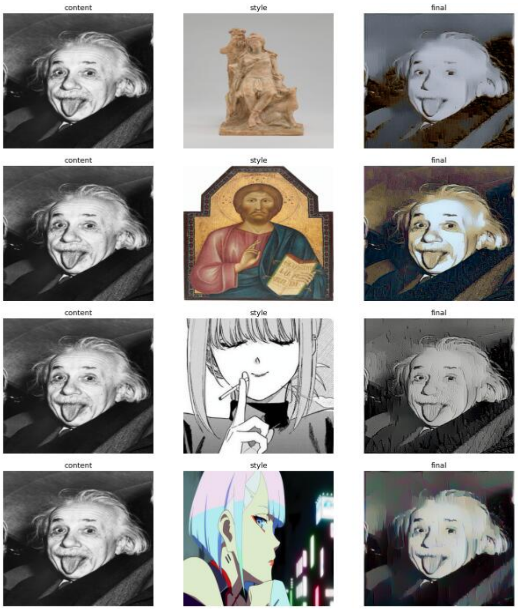


Figure 7: EfficientNet, complex background

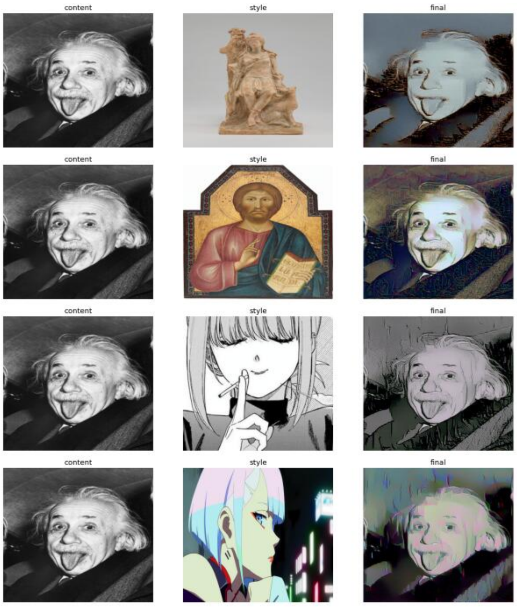


Figure 8: Mobile, complex background

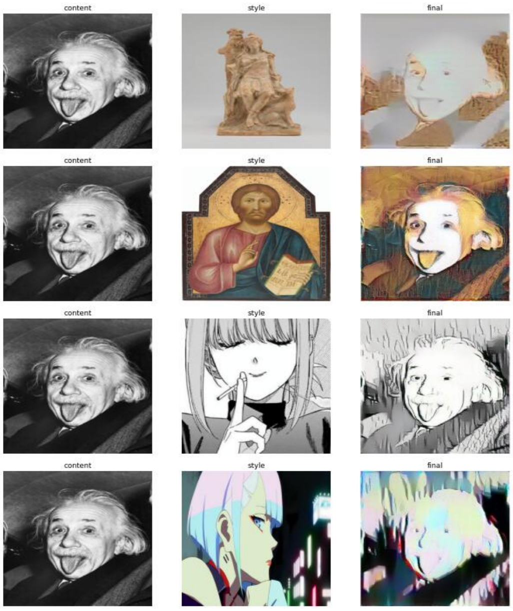


Figure 9: Inception-V3, complex background

7 Discussion

7.1 Performance Evaluation

EfficientNet's conservative style application underscores its strength in maintaining content integrity, a finding that is particularly valuable for applications where content preservation is crucial. Meanwhile, MobileNet offers a balanced approach, adeptly navigating between content and style without significantly favoring one over the other. VGG19's performance, however, stands out for its distinctive behavior to 'transfer' features in a manner akin to copying and pasting elements from the style image. This capability of VGG19 to translate style elements in a more literal sense is unique and provides a different aesthetic quality that may be desirable in certain artistic contexts.

7.2 Constraints and Future Exploration

The scope of our exploration was restricted due to time limitations and computational resources, preventing us from conducting extensive experimentation on the fine-tuning of layers, weights, and the intricate trade-off between content and style. We hypothesize that with a more in-depth and granular adjustment of these parameters, other models could potentially mimic the 'localization ability' that we observed in VGG19. This warrants further investigation to unlock and understand the full capabilities of each model.

Additionally, the evaluation metrics employed to gauge the performance of our NST models necessitate refinement. The current quantitative assessments primarily rely on loss functions which, while indicative of convergence, do not sufficiently encapsulate the artistic quality of the style transfer. The heterogeneity in the weights and architectures across different models renders a loss-based evaluation less informative. A more nuanced metric, such as ArtFID—proposed by Matthias Wright and Björn Ommer in 2022—promises a more sophisticated analysis by comparing the distribution of features in the stylized output to those in a dataset of artworks. Although preliminary attempts to implement ArtFID were thwarted by technical challenges and temporal constraints, its application presents a valuable avenue for future research. Properly leveraged, such a metric could provide pivotal insights into the aesthetic and technical efficacy of style transfer models, ultimately enhancing the objectivity and interpretability of NST performance evaluation.

7.3 Improvement Strategies

One promising approach to improve the quality of style-transferred images involves employing masking techniques to isolate specific areas of the image where the style transfer should either be highlighted or downplayed. By creating a mask for non-interest areas, we can potentially direct the model's attention and precision to areas where style transfer is most desired. This could help in mitigating the issue of complex backgrounds, allowing for more focused and coherent style applications on the content subject.
Another obvious way is to do an exhausting trial and error on the layers selection and their weights.

8 Ethical Considerations

The application of neural style transfer raises several ethical concerns, particularly the potential for creating deceptive images that could be used for misinformation. When our model applies the stylistic features of one image to the content of another, it creates a derivative work. The copyright status of this new creation can be complex. If the style image is copyrighted, the transformed image inherits this status, potentially infringing on the original creator's rights. Therefore, it is imperative to either use style images that are confirmed to be copyright-free or to obtain permission from the copyright holders before using the style transfer outputs for any commercial or public purpose.

References

[1] L. A. Gatys, A. S. Ecker, and M. Bethge, *A Neural Algorithm of Artistic Style*, arXiv preprint arXiv:1508.06576, 2016.

[2] M. Tan and Q. V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, In Proceedings of the 36th International Conference on Machine Learning, pp. 6105-6114, 2019.

[3] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*, arXiv preprint arXiv:1704.04861, 2017.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going Deeper with Convolutions*, In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1-9, 2015.

[5] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, *Rethinking the Inception Architecture for Computer Vision*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818-2826. doi: 10.1109/CVPR.2016.308.

[6] T. Ahmed and N. H. Sabab, ”Fig2 - Architecture of EfficientNet-B0 with MBConv as basic building blocks,” in *Classification and Understanding of Cloud Structures via Satellite Images with EfficientUNet*, Sep. 2020.

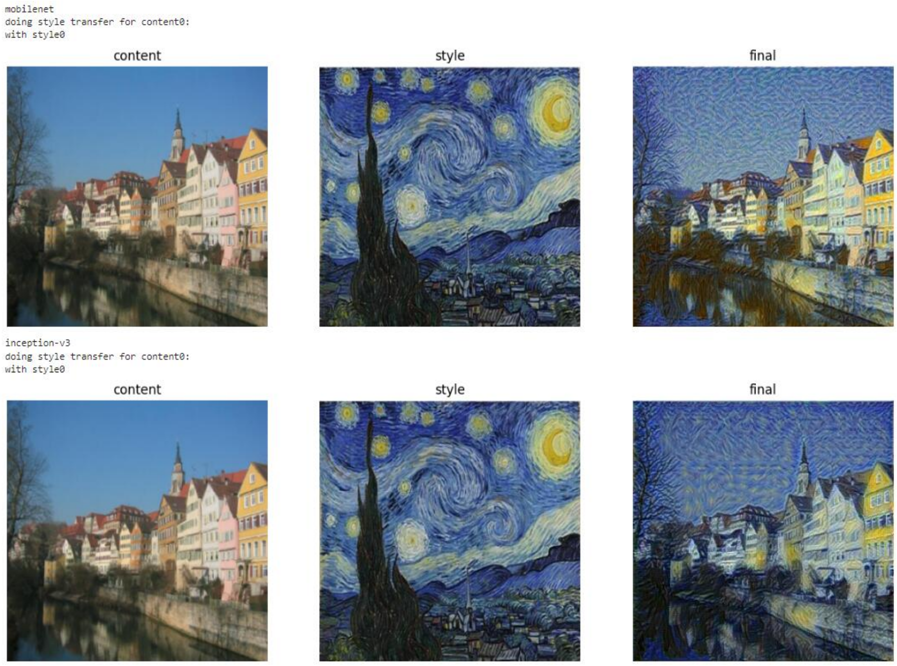
[7] Gaurav927. (2023). *Neural Style Transfer*. GitHub repository. Available at: `https://github.com/Gaurav927/Neural_Style_Transfer`

9 Appendix

9.1 Contribution

Elaine: Report, helped with NST,
Zheng Chen: NST for MobileNet and Inception-v3 and report,
Cathy: NST for EfficientNet and report
ChatGPT: Helped in coding and report

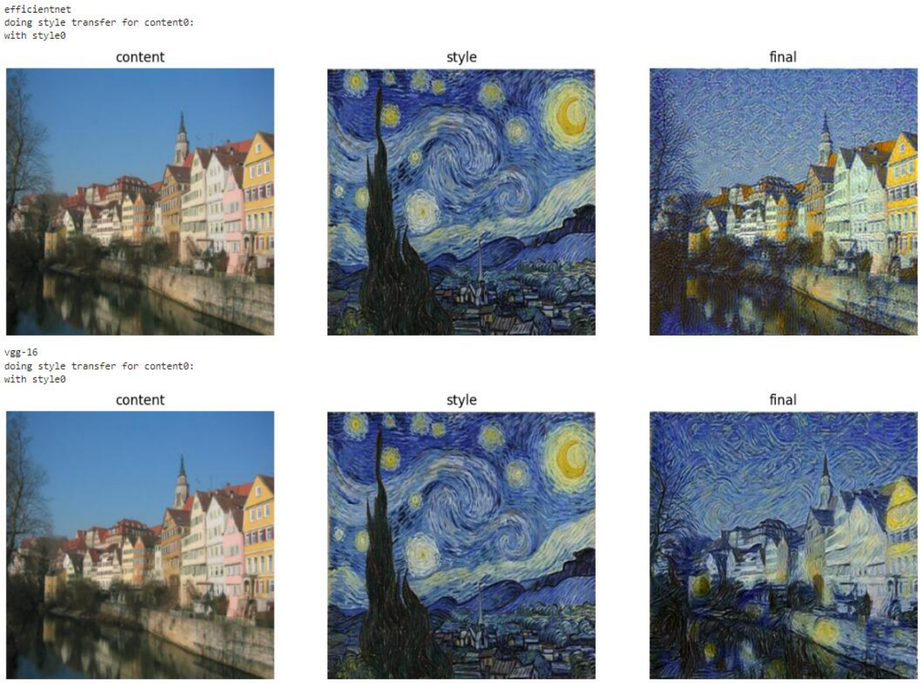
9.2 More Output



```
mobilenet
doing style transfer for content0:
with style0
```

```
inception-v3
doing style transfer for content0:
with style0
```

Figure 10: MobileNet, Inception-V3 output



```
efficientnet
doing style transfer for content0:
with style0
```

```
vgg-16
doing style transfer for content0:
with style0
```

Figure 11: EfficientNet, VGG-19 output

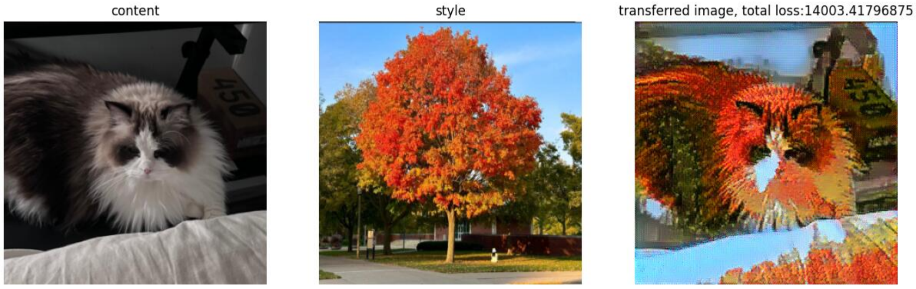


Figure 12: An artistic cat generated by our NST on Inception