Extensions to Neural Style Transfer

Team 4D Tensor
Praveen Iyer, Swarnita Venkatraman, Sneha Bandi, Neha Chawla

Abstract

Given a content and style image, the process of applying styles to generate an output image that retains the core elements of the content image but appears to be in the style of the reference image is known as Neural Style Transfer (NST) (Gatys et al., 2015a). In recent times, NST has achieved good performance, but studies highlighting improvements in stability, quality, flexibility, and evaluation are still in its early stages, specifically using user feedback. To build upon these challenges, we propose an extension to the original NST technique to enhance the final style of the output image, using the novel pipeline of Cascade Style Transfer (CST). Our key idea focuses on using CST to combine multiple style images, alongside novel quantitative and qualitative evaluation methods to assess the stylistic quality of generated image. Our cascade framework involves using multiple architectures in serial based on traditional Gram Matrices, finally evaluating the stylized output using three quantifiable factors - Content Fidelity (CF), Global Effects (GE) and Local Patterns (LP) and qualitative evaluation.

1 Introduction

Multiple Style Transfer (MST) in Neural Style Transfer (NST) enables styles of multiple images to be transferred to a single content image by modification of style losses. Multi-Objective network (MO) uses multiple style images to optimize multiple objectives, for content preservation and style transfer, simultaneously while Cascade style transfer (CST) is a neural style transfer method that iteratively applies style to a content image to preserve both style and original content, sequentially.

Existing challenges with multiple style transfer include increased training time due to a separate network tied to one style. Adding multiple styles in MO can lessen the impact of each individual style, especially when the styles are somewhat different (Makow and Hernandez, 2017). Higher style loss exists when blending two images compared to having just one style image (Makow and Hernandez, 2017). On the other hand, CST is shown to improve the stylistic quality and flexibility of the generated image. (Wang et al., 2021b)(Wang et al., 2020) Style from the style image is known to be preserved more in CST, while in traditional MOs, content from the content image is known to be preserved more. (Wang et al., 2021b) Exploring possibilities of performing multiple style transfer using CST to exploit its benefit of preserving style gives us room to incorporate multiple styles. Besides, existing literature on using CST for Multiple Style Transfer has not been found yet. Two style images might have contradicting styles that can lead to unsatisfactory output. The ordering of style images in the CST will also matter leading to vastly different outputs for different ordering of input. There is no ground truth for NST.

Based on the complementary strengths that CST has when compared to the challenges of Multiple Style Transfer using MO, the **hypothesis** of the project was formulated. Hypothesis of the project is, "Using CST for Multiple Style Transfer enables generation of images that incorporate multiple styles with better quality and style preservation"

Evaluating and improving the stylization quality remain two important open challenges, standard quantifiable metric implementation to measure NST performance does not exist. This means that for evaluation we rely on qualitative methods which can be challenging. This can lead to subjective opinions mattering more rather than having a clear objective which is being optimized. Designing more fine grained metrics might also be needed to improve the quality of results.

We evaluate the hypothesis [1] to understand if the style is known to be preserved more in CST than

[1] https://github.com/praveen-iyer/CSCI566-Project

in traditional MO to enable generation of images with better quality. For testing the above hypothesis we use evaluation metrics Content Fidelity, Global Effects, and Local Patterns along with a qualitative user study.

2 Related Work

In recent times, though NST has achieved good performance, studies highlighting improving its stability, quality, flexibility, and evaluation are still in its early stages. There is a dearth of literature on quantitative metrics for evaluating the quality of generated images through NST, especially when using CST with multiple styles.

Traditional Convolutional Neural Networks (CNN) hierarchically stack to form content representations. It is possible to form style representations by capturing the correlations of different filter responses within each layer. In (Gatys et al., 2015a) produced the first work investigating the extraction style representations from images. Additionally, much work has been done on exploring the perceptual possibilities of neural style, where novel extensions to the original neural style algorithm have greatly improved the perceptual quality of generated images, as well as introduced a number of new features. These features include multiple style transfer (Dumoulin et al., 2017), color-preserving style transfer (Gatys et al., 2016), as well as content-aware style transfer (Gatys et al., 2017).

To allow for the styles of multiple images to be transferred to a single content image, simple modifications to the style loss function have been proposed in this paper (Makow and Hernandez, 2017). The paper explains two perceptual loss functions of interest: feature reconstruction loss and the style reconstruction loss. Both make use of a pre-trained loss network trained for image classification on the ImageNet dataset. For Feature Reconstruction Loss, the pixels of the output image have similar feature representations as computed by the loss network, and in Style Reconstruction Loss, the style reconstruction loss penalizes differences in style, such as differences in colors and textures (Gatys et al., 2015b), common patterns, etc. This loss formulation selects the style layers and weights independently for each style image enabling the model to weigh overall effects of each style image on the generated image.

The paper (Wang et al., 2020) focuses on leveraging CST by combining existing approaches in Serial Style Transfer (SST) architecture to improve quality and Parallel Style Transfer (PST) architecture to improve quality and flexibility. It is one of the first papers to focus on domain independent style transfer across three domains namely artistic, semantic, and photo realistic using one content and one style image. For SST architecture, it proposes a linear combination of n methods chosen thoughtfully based on each of their strengths with regards to factors like content fidelity, global color, and local patterns with performance levels of High, Medium, and Low. The paper suggests reasonably concluded serial schemes for each of the domains based on their end goal. In PST architecture, a single backbone is shared by all methods and the total loss is optimized in parallel. The proposed ParallelNet architecture uses weighted losses and by increasing the weights of respective losses both global color and local textures of the style image of two different methods can be achieved. A user study was conducted by randomizing various combinations of style and content images where users were shown results of 10 compared methods and SST/PST outperformed the state of the art models for all three domains. Besides, the paper also mentions more effective and efficient scheme designs.

In this paper (Wang et al., 2021b), three new quantitative factors (Content Fidelity, Global Effects, and Local Patterns) for evaluating NST are introduced. Using these factors ideas have been touched upon which involve the usage of CST and user feedback. It brings up the point where CST as an architecture is biased more to style transfer than content preservation. This can be an important feature in the case of multi style transfer, which we aim to do. Additionally, the paper also discussed how the user feedback can be incorporated using the three quantitative evaluation factors by modifying the loss function to include the user preferences. This paper also provides a thorough comparison of traditional multi-objective networks and CST architectures with average stylization ranking metrics clearly highlighting the superiority of CST over MO in this aspect.

"Interactive Artistic Multi-style Transfer" (Wang et al., 2021a) is a research paper that proposes a novel approach for interactive multi-style transfer in artistic images. A new system is introduced that allows users to manipulate the style and content of an input image by selecting different style images

and adjusting various parameters. Their method combines feature normalization, multi-layer transformation, and iterative optimization to achieve high-quality style transfer while maintaining the content of the original image. The proposed approach enables users to create customized artistic images with multiple styles, which can be used for a variety of applications such as graphic design, digital art, and visual communication.

3 Implementation Details

For evaluating the hypothesis, three different model architectures were combined in SST. Each generated result was then compared to a baseline MO model to check if using CST improved our results as hypothesized. For evaluating the performance, three quantitative evaluation metrics were implemented from scratch. User-based qualitative evaluation was also designed to support it.

3.1 Dataset

Since NST does not fall under supervised learning or unsupervised learning, there is no fixed dataset that we are using. For testing and development purposes we have been using the same set of images as provided in the GitHub repository (Gordić, 2020).

3.2 Baseline Model

As a starting point, we decided to modify the traditional Gram Matrix network as a MO which will be able to handle our multi style image setting. This was giving satisfactory results visually and was used as a baseline model, shown in Fig 1.

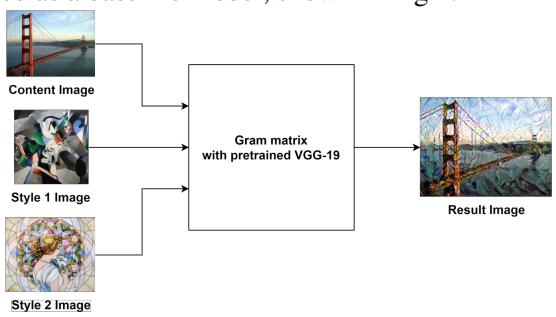


Figure 1: Baseline for Multiple Style Transfer using MO model

As a part of the baseline model (Gatys et al., 2015a), style transfer is performed, using gradient descent on random noise to minimize the deviation of target image from content. The deviation from the style representation of the style image, is defined as a set of Gram correlation matrices G_l with the entries $G^l_{ij} = (F^l_i, F^l_j)$,where l is the network layer, i and j are two filters of the layer l and F^l_k is the array (indexed by spatial coordinates) of neuron responses of filter k in layer l. We have therefore used the VGG-19 network as backbone. It performs better at style transfer, as its architecture captures more information compared to other networks like AlexNet, GoogLeNet which strongly compress the input at first convolutional layer, thus losing a lot of fine details. L-BFGS is preferred over Adam due to its success as a second-order optimization algorithm that can handle larger batch sizes.

3.3 Models Explored

SST is a technique under CST that involves applying a sequence of style transfer algorithms or models to an image, with each subsequent model building on the output of the previous one. SST was implemented with two different architectures connected serially one after the other.

Many models were considered for exploration based on the paper (Wang et al., 2020). Three different models were used as architecture 1 while the architecture 2 was kept constant to be the Gram Matrix model as shown in Fig 2. A summary of each model used is given below. The content image and style 1 image are given to architecture 1. The output of architecture 1 becomes the content image for architecture 2 alongside the style 2 image and the output images are generated.

3.3.1 Model 1: Gram Matrix with pretrained VGG-19

The traditional Gram Matrix model with pretrained VGG-19 backbone (as seen in the baseline) was connected serially two times, as both architecture 1 and 2. The stylistic characteristics of style 1 were replaced by the stylistic characteristics of style 2 on the content image. Results obtained were not promising.

3.3.2 Model 2: Variational AutoEncoder (VAE)

A VAE model was used as architecture 1 along with the gram matrix model as architecture 2. ST-VAE (Style Transfer VAE), is a VAE for latent space-based style transfer which performs multiple style transfer by projecting nonlinear styles to a linear latent space, enabling to merge styles via linear interpolation before transferring the new style to the content image (Liu et al., 2021). According to the paper this method is faster and flexible for multiple style transfer compared to baseline methods. The ST-VAE method was implemented using exist-

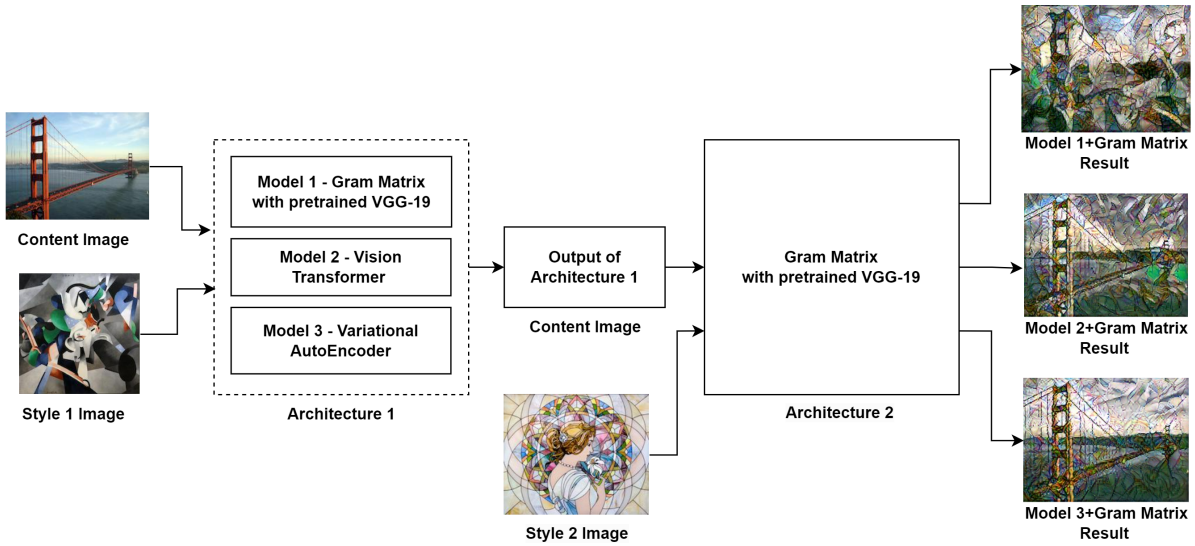


Figure 2: Three different models explored as Architecture 1 in SST

ing code for both single and multiple style transfer. For multiple style transfer the existing code base focuses on using four style images which was modified according to the project goals which aims at combining two style images. It showed promising results when implemented separately and hence was chosen to be further investigated in SST. The results obtained looked promising when compared to the gram matrix model connected serially twice.

3.3.3 Model 3: STyle TRansformer (STTR)

A Vision Transformer model was used as architecture 1 along with the gram matrix model as architecture 2. Existing approach adopts a global feature transformation to transfer style patterns into content images. Such a design usually destroys the spatial information of the input images and fails to transfer fine-grained style patterns into style transfer results. To solve this problem, we use the STTR network proposed in (Wang et al., 2022) which breaks both content and style images into visual tokens to achieve a fine-grained style transformation based on attention mechanism (Deng et al., 2021). When implemented for single style transfer, it showed good results. Hence it was used with gram matrix in serial to investigate its properties for combining multiple style images, in SST.

4 Evaluation

4.1 Quantitative Evaluation

We implemented three metrics (Wang et al., 2021b) from scratch to incorporate them into the existing code base for result evaluation. Creating an efficient implementation for local patterns was particularly challenging as it involved generating and comparing all pairs of patches in two images.

4.1.1 Metric 1: Content Fidelity (CF)

It is a way to measure the extent to which the content of the original image is preserved in the stylized image. Since NST involves transforming the content of the original image to match the style of a style image, it is important to evaluate the quality of the output with respect to how much it retains the original content. In the equation below, $\overrightarrow{c}$ represents the content image while $\overrightarrow{x}$ represents the stylized result. N represents the total number of layers and $f_l(.)$ represents the deep feature activations that are extracted as we go through each layer from l=1 to N. This results in the content fidelity factor.

$$CF\,(\overrightarrow{x}, \overrightarrow{c}) = \frac{1}{N} \cdot \sum_{l=1}^{N} \frac{f_l\,(\overrightarrow{x})\,.f_l\,(\overrightarrow{c})}{\|f_l\,(\overrightarrow{x})\|\,.\|f_l\,(\overrightarrow{c})\|}$$

4.1.2 Metric 2: Global Effects (GE)

It is a measure of overall visual quality in NST that evaluates the similarity between the transferred style and the style image and it reflects the initial impression that the image leaves on human visual perception. It encompasses two aspects: Global Colors (GC) and Holistic Textures (HT), which were identified as important factors in user studies. GC compares the similarity of color histograms, while HT uses the Gram matrix to capture the HT across multiple layers. The GE factor is the average of both GC and HT. In the equation below,

$\vec{s}$ represents the style image and $\vec{x}$ represents the stylized result. c represents the channel and $hist_c(.)$ represents the color histogram vector for the c[th] channel. This results in the global colors factor.

$$GC\left(\vec{x},\vec{s}\right)=\frac{1}{3}.\sum_{c=1}^{3}\frac{hist_c\left(\vec{x}\right).hist_c\left(\vec{s}\right)}{\left\|hist_c\left(\vec{x}\right)\right\|.\left\|hist_c\left(\vec{s}\right)\right\|}$$

Then, in the next equation below, $G(.)$ represents the gram matrix calculation and this equation results in the HT.

$$HT\left(\vec{x},\vec{s}\right)=\frac{1}{N}.\sum_{l=1}^{N}\frac{G\left(f_l\left(\vec{x}\right)\right).G\left(f_l\left(\vec{s}\right)\right)}{\left\|G\left(f_l\left(\vec{x}\right)\right)\right\|.\left\|G\left(f_l\left(\vec{s}\right)\right)\right\|}$$

Then, in the final equation below, both the GC and HT factors are taken into account to compute GE.

$$GE\left(\vec{x},\vec{s}\right)=\frac{1}{2}.\left(GC\left(\vec{x},\vec{s}\right)+HT\left(\vec{x},\vec{s}\right)\right)$$

4.1.3 Metric 3: Local patterns (LP)

It is an evaluation metric to measure the extent to which patterns like brush strokes and exquisite motifs from the style image are transferred to the stylized image. This is done by using two components: the first assesses the similarity between the local patterns in the original and stylized images, while the second compares the variety of the pattern categories identified in both images.
To calculate this we first extract 3x3 patches from $f_l\left(\vec{x}\right)$ and $f_l\left(\vec{s}\right)$ denoted by $\left\{\phi_i^l(\vec{x}))\right\}_{i\in n_x}$ and $\left\{\phi_j^l(\vec{s}))\right\}_{j\in n_s}$ where n_x and n_s are the number of patches respectively.

$$CM\left(i\right):=\underset{j=1\ldots,n_s}{argmax}\frac{\phi_i^l\left(\vec{x}\right).\phi_j^l\left(\vec{s}\right)}{\left\|\phi_i^l\left(\vec{x}\right)\right\|.\left\|\phi_j^l\left(\vec{s}\right)\right\|}$$

$$LP_1\left(\vec{x},\vec{s}\right)=\sum_{l=1}^{N}\sum_{i=1}^{n_x}\frac{\phi_i^l\left(\vec{x}\right).\phi_{CM(i)}^l\left(\vec{s}\right)}{\left\|\phi_i^l\left(\vec{x}\right)\right\|.\left\|\phi_{CM(i)}^l\left(\vec{s}\right)\right\|}$$

$$LP_2\left(\vec{x},\vec{s}\right)=\frac{1}{N}\sum_{l=1}^{N}\frac{t_{cm}^{'}}{t_s^{'}}$$

Here $t_{cm}^{'}$ and $t_s^{'}$ denote the number of unique patches in $\left\{\phi_{CM(i)}^l(\vec{s}))\right\}$ and $\left\{\phi_j^l(\vec{s}))\right\}$ respectively.

$$LP\left(\vec{x},\vec{s}\right)=\frac{1}{2}.\left(LP_1\left(\vec{x},\vec{s}\right)+LP_2\left(\vec{x},\vec{s}\right)\right)$$

4.2 Qualitative User Evaluation

To conduct a qualitative analysis, a Google form with 4 questions was created based on TA feedback. The form consisted of four questions, with one for each metric and one for the overall quality. We received over 30 responses for this exercise from technical individuals working in the field of Computer Science. The questions asked to the user were:

1. Which of the resulting images above is able to preserve the original content image? (CF)
2. Which result image has a similar color palette and textures compared to the two style images shown above? (GE)
3. Which resulting image has more lines/edges/patterns from the two style images shown above? (LP)
4. Which image do you feel is overall able to contain both the styles and visually pleasing? (Overall Quality)

We permitted users to choose multiple responses for the first question related to Content Fidelity, as our initial assumption was that the variables were independent in both architectures. For the other questions, we opted for a single option as our aim was to determine the relative ordering of output images for each metric.

4.3 Observations

The output image was generated by combining content image with two style images using different models. Fig 3 shows the results from different model combinations.

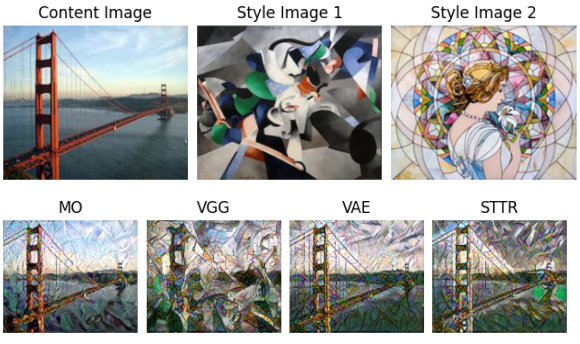


Figure 3: Content image: 'Golden Gate', style 1 image: 'Udnie', style 2 image: 'Mosaic'

Based on results in Fig 3 different observations were made after running evaluation metrics, listed in table 1 and 2.

Q	Survey Questions	MO(%)	VGG(%)	VAE(%)	STTR(%)
1.	Which of the resulting images above is able to preserve the original Content image?	100.00	0.0	100.0	40.00
2.	Which result image has a similar color palette and textures compared to the two style images?	26.67	46.67	6.67	20.00
3.	Which resulting image has more lines/edges/patterns from the two style images?	0.00	46.67	13.33	40.00
4.	Which image do you feel is overall able to contain both the styles and visually pleasing?	33.33	0.00	46.67	20.00

Table 1: Qualitative observations of MO vs CST
A parallel user study was performed against each quantitative metric (CF, GE, LP), to evaluate quality of output image. The above table denotes observations of the percentage distribution of users for each question asked in survey.

Metrics		MO	VGG	VAE	STTR
CF		0.38	0.26	0.82	0.51
GE	style 1	0.73	0.78	0.76	0.72
	style 2	0.81	0.84	0.81	0.86
LP	style 1	0.55	0.51	0.49	0.42
	style 2	0.53	0.59	0.43	0.49

Table 2: Quantitative metric observations MO vs CST.
The table shows, different values calculated for the three quantitative metrics (CF, GE, LP). GE and LP is calculated for each style image.

Table 1 shows observation about the human study. For each question asked, larger percentages were picked towards Cascaded Networks - VGG, VAE, STTR. For question 1, since users were enabled multiple selections, they picked MO and CST outputs equivalently. For the second question, higher or equivalent percentages were picked by the user for CST. Intriguingly in the third question for LP, nearly no one picked MO and favored CST majorly. Finally the overall winner observed was CST, following our hypothesis over MO.
Further **Table 2** depicts inference about the three metrics that CF increased, GE slightly moved in both the directions, such that it decreased for style 1 and increased for style 2 and LP decreased for both the styles.

5 Results and Discussion

Table 1 and Table 2 depict the results, and based on them, we discuss our initial hypothesis and the findings in the Table shown in **Fig 4**.
In Fig 4 the Blue area denotes MO and rest is a part of different Cascaded architectures. The pie charts denote the user evaluation (outer circle) and quantitative metrics (inner circle) projected against each other. We can clearly see non-blue areas are higher on the outer circle as well as cogent to the inner circle. One of the significant results was that none of the users picked MO output in question 3 (Table 2) that is for local patterns. For GE there was even distribution, when comparing MO and CST. Overall the users picked CST outputs more than MO.
Originally MST architecture was implemented by jointly minimizing the feature loss and the style loss formulated as the difference of Gram matrices. As a part of the alternate architectures, VAE transfers textures and styles efficiently and also successfully preserves the content information during style transition (Liu et al., 2021), which we could see in our results as well. Transformer network (STTR) preserves spatial information of the content structure and style patterns as it matches the style tokens onto content tokens in a fine-grained manner, which is coherent with our results.

6 Conclusion

Based on our results, we summarized a few conclusions. Content was equivalently preserved in CST as in MO. Style qualities like color palette and texture were significantly preserved in CST. Style patterns like lines, edges and patterns were more preserved in CST, true to our hypothesis. Based on the different architecture combination results we can conclude that CST results seem promising and can be considered to be a viable alternative to MO for better quality and style preservation in the generated image. Additionally we also show that quantitative metrics can be a feasible and reproducible option that can be used for evaluating NST performance in line with the prevalent qualitative evaluation techniques.

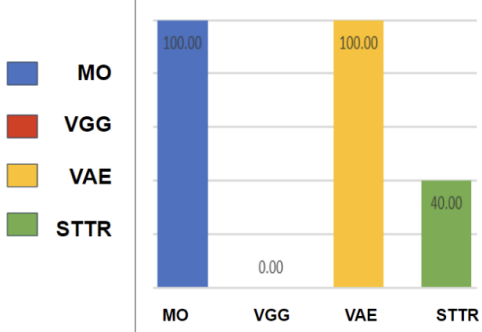
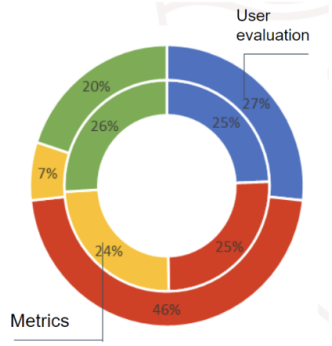
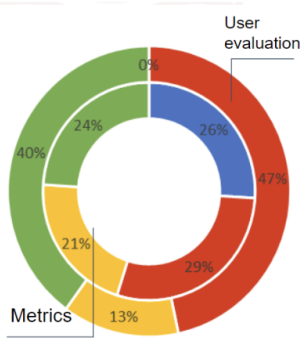
Metrics	Content Fidelity	Global Effects	Local Patterns
Hypothesis	CF_{CST} ,CF_{MO} are independent	$GE_{CST} > GE_{MO}$	$LP_{CST} > LP_{MO}$
Result	Coherent to hypothesis, Increases, mutually exclusive, within the bounds, Preserves content fairly	Slightly varied in both directions, i.e mixed distribution, matching color palette-texture expectations	Decreases, Style parameters - lines and edges are better preserved, supports our hypothesis
			

Figure 4: Evaluation Hypothesis and Results
Table shown above denotes the comparison between our initial hypothesis and our deductions against the same, based on our observations for each metric, both quantitatively and qualitatively. It also depicts the parallel conclusion between both kinds of evaluations (quantitative and qualitative)

7 Future Scope

The future scope of our work involves further improvements to our proposed CST method. One extension consists of examining alternate architecture approaches, such as PST to extend our CST technique. Additionally, the original loss function can be investigated to extend its capability to transfer multiple style images optimally. Furthermore, enhancement based on user feedback can be explored by streamlining it to existing neural style models. User evaluations can be mapped to current metrics for adjusting different parameters of an image. Lastly, the incorporation of visual parameters such as brightness, contrast, hue, etc. can be considered for real-time rendering to further improve the style transfer process.

Acknowledgements

We thank **Bingjie Tang** for providing valuable feedback and support throughout our project. We thank Professor **Jesse Thomason** for providing the opportunity to pursue our research. Lastly, we thank Google Cloud Platform which provided us with vital computational resources that helped us accomplish our project goals.

References

Yingying Deng, Fan Tang, Xingjia Pan, Weiming Dong, ChongyangMa, and Changsheng Xu. 2021. Stytr2̂: Unbiased image style transfer with transformers. *ArXiv*, abs/2105.14576.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In *International Conference on Learning Representations*.

Leon A. Gatys, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. 2016. Preserving color in neural artistic style transfer. *CoRR*, abs/1606.05897.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015a. A neural algorithm of artistic style. *Journal of Vision*, abs/1508.06576.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. 2015b. Texture synthesis using convolutional neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 262–270, Cambridge, MA, USA. MIT Press.

Leon A. Gatys, Alexander S. Ecker, Matthias Bethge, Aaron Hertzmann, and Eli Shechtman. 2017. Controlling perceptual factors in neural style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3730–3738.

Aleksa Gordić. 2020. Reference GitHub repository. https://github.com/gordicaleksa/pytorch-neural-style-transfer.

Zhi-Song Liu, Vicky Kalogeiton, and Marie-Paule Cani. 2021. Multiple style transfer via variational autoencoder. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2413–2417.

Noah Makow and Pablo Hernandez. 2017. Exploring style transfer: Extensions to neural style transfer. *Stanford University Report*.

Jianbo Wang, Huan Yang, Jianlong Fu, T. Yamasaki, and Baining Guo. 2022. Fine-grained image style transfer with visual transformers. In *Asian Conference on Computer Vision*.

Xiaohui Wang, Yiran Lyu, Junfeng Huang, Ziying Wang, and Jingyan Qin. 2021a. Interactive artistic multi-style transfer. *International Journal of Computational Intelligence Systems*, 14.

Zhizhong Wang, Lei Zhao, Haibo Chen, Zhiwen Zuo, Ailin Li, Wei Xing, and Dongming Lu. 2021b. Evaluate and improve the quality of neural style transfer. *Computer Vision and Image Understanding*, 207:103203.

Zhizhong Wang, Lei Zhao, Qihang Mo, Sihuan Lin, Zhiwen Zuo, Wei Xing, and Dongming Lu. 2020. Cascade style transfer. *Under review as a conference paper at ICLR 2020*.

A Division of Labor

Each team member played a pivotal role in ensuring the successful completion of the project. The division of work was based on individual strengths and interests, with regular discussions to ensure equal distribution of responsibilities. The team collaborated smoothly using GitHub as the version control technology.

The **project proposal** was meticulously planned, with:

- **Sneha** focusing on the introduction and outlining the objective and goals,
- **Swarnita** identifying existing work and key ideas for novelty,
- **Neha** highlighting the potential impact of the project; and
- **Praveen** identifying potential challenges and solutions.

The team collectively estimated the computation costs, divided the work, and established a timeline.

During the **project pitch**, all four members collaborated to prepare the slides and content, with Sneha delivering the pitch.

In the **survey report**, each team member contributed to the Related Work section by reading and summarizing one research paper.

- **Praveen** worked on quantitative evaluation related work,
- **Swarnita** worked on cascade style transfer architecture research,
- **Sneha** worked on multiple style techniques; and
- **Neha** worked on exploration of interactive styling.

Each team member also contributed two challenges to the 'Challenges' section that were identified from the respective papers and earlier from the proposal. All team members worked on contributing a set of references to the survey report and the final addition to Overleaf Latex format.

Before the **midterm report**:

- Implementation of the three evaluation metrics was a collaborative effort of **Swarnita, Neha and Praveen**,
- **Praveen** worked on the Gram matrix-based method,
- **Sneha** implemented the Vision Transformer-based method; and
- **Swarnita** handled the VAE-based method.

Before the **final report**:

- **Praveen** handled the final efficient implementation and verification of the three evaluation metrics,
- **Neha** designed the qualitative study,
- **Swarnita** integrated the VAE-based method in a SST setting; and
- **Sneha** integrated the Transformer-based architecture in a SST setting and interpreted observations from the different architecture combinations and qualitative study.

All team members collaborated in gathering responses for the study. Finally, for the final presentation and report, all team members worked together and contributed equally.

B Appendix

Fig 5, shows the transition of output image alongside style addition, simultaneously from both the input styles. We can see different levels or percentage effect of styles, for MO. Similarly Fig 6, shows the transition of output image alongside style 1 addition, and Fig 7 shows style 2 addition transitions.

Figure 5: Multi-objective Network Transitions at various iterations for only style 1 addition

Figure 6: Gram Matrix + Gram Matrix SST transitions at various iterations for style 1

Figure 7: Gram Matrix + Gram Matrix SST transitions at various iterations for style 2

Based on the inputs in Fig 8, left side of Fig 9 depicts visually satisfactory results for MO and the right side shows style 2 dominating over style 1. This results from using the same architectures for SST.

Figure 8: Content image:'Figures', style 1 image:'Giger Crop',style 2 image: 'Mosaic'

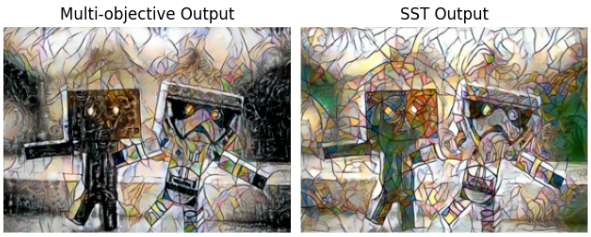


Figure 9: MO Output and Gram Matrix + Gram Matrix SST Output

For further exploration with different models following were the results upon which we built the cascaded networks i.e. combinations of different architectures. Implementation of just the VAE model separately for MST is shown in Fig 10. The obtained results on combining Transformer Network serially two times for MST are shown in Fig 11. For additional exploration we also changed the order of the style incorporated in the final image to study its effect, as shown in Fig 12.

Figure 10: VAE Output with Style Image 'Ben Giles' then 'Starry Night'

Figure 11: STTR Output with Style Image 'Udnie' then 'Mosaic'

Figure 12: STTR Output with Style Image 'Mosaic' then 'Udnie'

Fig 13, 14, 15, 16 and 17 are the screenshots showing the Google form created for the user study.

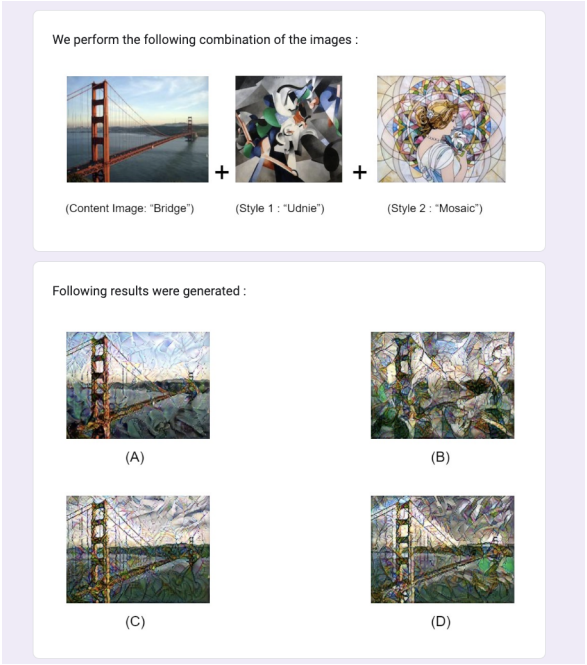


Figure 13: Qualitative User Study Form Image Options

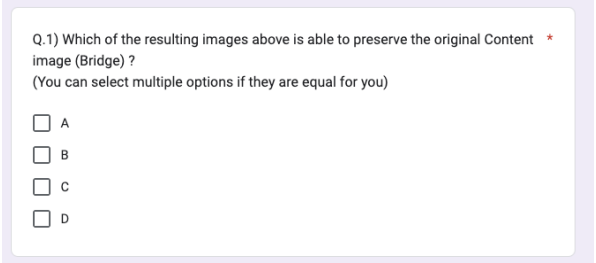


Figure 14: Qualitative User Study Form Question 1

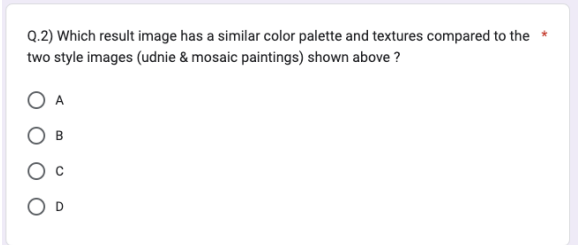


Figure 15: Qualitative User Study Form Question 2

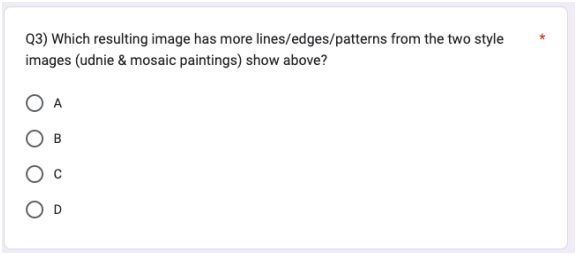


Figure 16: Qualitative User Study Form Question 3

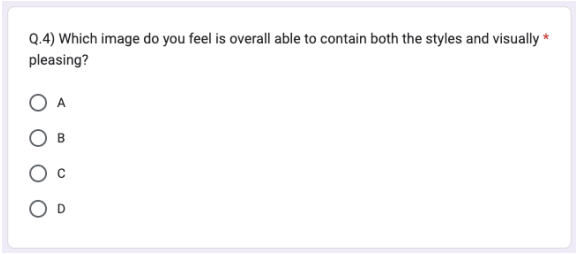


Figure 17: Qualitative User Study Form Question 4