TEXT-BASED IMAGE STYLE TRANSFER

Members:

Anushka Jain:21ucc022

Elish Baraiya:21ucs077

Harsha Rani:21ucs088

Varad Bane:21uccl11

WHAT IS TEXT-BASED STYLE TRANSFER

- Style transfer methods require reference style images to transfer texture information of style images to content images. However, in many practical situations, users may not have reference style images but still be interested in transferring styles by just imagining them.

CONTENT IMAGE

-------->

TEXT INPUT:
OIL PAINTINGS OF FLOWER

STYLED IMAGE

PROBLEM STATEMENT

- In the realm of text-based image style transfer, numerous research papers propose innovative methodologies aiming to generate visually appealing images by transferring artistic styles from textual descriptions. However, despite advancements in this field, existing approaches often exhibit limitations in terms of image quality, style fidelity, and computational efficiency. Our project seeks to address these limitations by critically reviewing seminal papers, selecting a baseline model as a point of reference, and identifying key shortcomings.

LITERATURE REWIEW

- Link: https://drive.google.com/file/d/1K9qxsDNUvrM6OfwfIMWSUB_SRaU_XGHc/view?usp=sharing

BACKGROUND

- **Image style transfer:**

Neural Style Transfer : By Gatys.et.al

Loss function: $\mathcal{L}_{\text{total}} = \sum_{l \in [L]} \alpha_l \mathcal{L}'_C + \gamma \sum_{l \in [L]} \beta_l \mathcal{L}'_S$

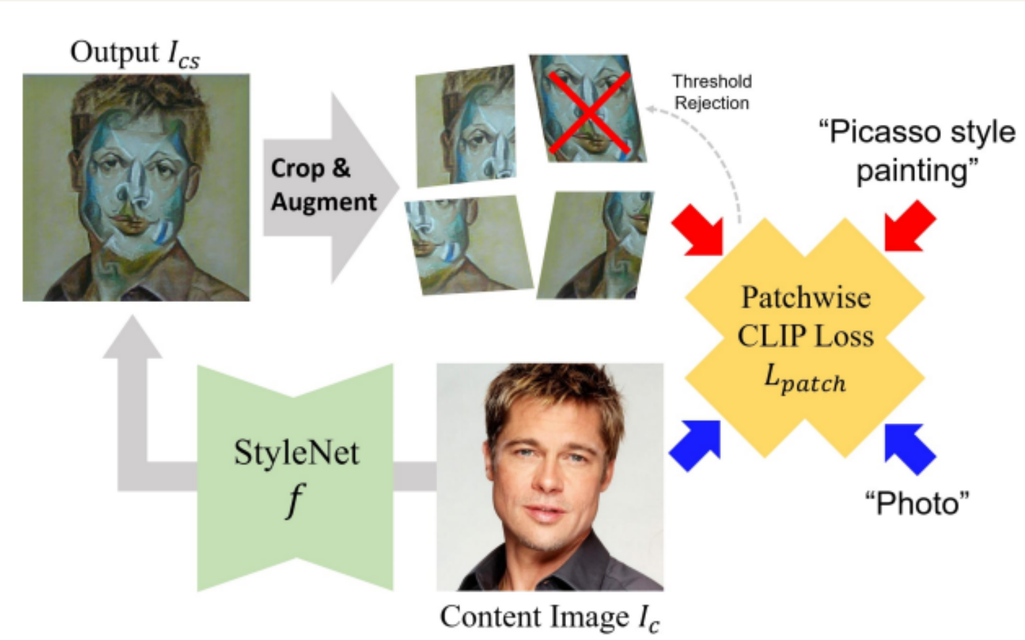
Each is a MSE Loss

- **Content Loss between the Content image feature map and stylized image feature map from same layer of VGG19**
- **Style (obtained from Gram Matrix) of style image with style of stylized Image.**
- **Learning Mechanism: Pixel Optimization of output stylized image through back propagation**

Text Based Image Style Transfer :

- CLIP (Contrastive Language Image Pre-training)
- CLIPStyler (Kwon.et.al)

Methodology:



Loss function:

$$L_{\text{total}} = \lambda_d L_{dir} + \lambda_p L_{\text{patch}} + \lambda_c L_c + \lambda_{tv} L_{tv}$$

BASELINE PAPERS

Spectral Clip

Directly using CLIP for style can result in unwanted artifacts. Spectral CLIP fixes this by using CLIP's spectral representation. It blocks out frequencies with common artifacts, ensuring the generated images match the desired style without those unwanted elements.

Multimodality-guided Image Style Transfer

This paper improves upon traditional approaches by allowing for flexible style specification and achieves state-of-the-art results in both Text-guided Image Style Transfer (TIST) and MultiModality-guided tasks.

SPECTRAL CLIP

- SpectralCLIP prevent both textual and visual artifacts in CLIP-guided style transfer.
- It include representation of images using CLIP encoders,frequency analysis and the application of band stop filter for artifact seperation Methods used:

1. Spectral based filtering
2. Computing an image-text similarity
3. Band selection

“pop art”

content
image

CLIPStyler
output

SPECTRAL FILTERING

1. A CLIP vision encoder(here ViT) represent the content image as a grid of vectors(one for each patch)
2. encoder also includes a class token, we get n = 1+k*k vectors, which we flatten into a sequence: V = Ev(I), where V = {v0, ..., vi , ..., vn-1}.
3. spectral representation of V can be obtained.
4. The corresponding frequency domain coefficients can be obtained though DCT.
5. specific frequency bands, likely corresponding to artifacts, are filtered out using a band-stop filter.
6. This filtered representation is then back-projected to the original CLIP space, providing a refined image representation that can be utilized for image-text similarity computation while disregarding artifact-related frequencies.

COMPUTING AN IMAGE-TEXT SIMILARITY

This filtered CLIP embedding of the stylized image (generated by CLIPStyler) can then be used exactly in place of earlier CLIP embedding before spectral analysis.

BAND SELECTION

- To simplify frequency filtering fixed frequency bands are defined(here 5) categorized by their corresponding periods.
- These bands are chosen based on the typical scales of artifacts in images.
- Three effective filtering strategies (c1, c2, c3) are identified, each targeting artifacts at different scales.
- The best-performing strategy is selected through visual inspection and applied consistently for each style during image generation, ensuring artifact prevention without the need for individualized filter selection.

REPRODUCED RESULT OF PAPER

------->

TEXT INPUT:
OUTSIDER ART

MOTIVATION

An attempt to Improve the results

Limitation:

1. **This work defines three general band combinations that effectively produce cleaner stylized images.**
2. **empirically analyzed artifact patterns present in a range of artistic styles therefore could only give handcrafted filters for certain styles.**

Possible Solution: A more promising alternative for future work is to automatically select frequency bands that cater to a target style

ABSTRACT OF MMIST

- The research paper focuses on a specific approach to achieving MMIST. It might propose a novel method that builds upon existing techniques like CLIPStyler (which leverages pre-trained models for text-image understanding) but offers improvements in areas like style representation or content preservation.

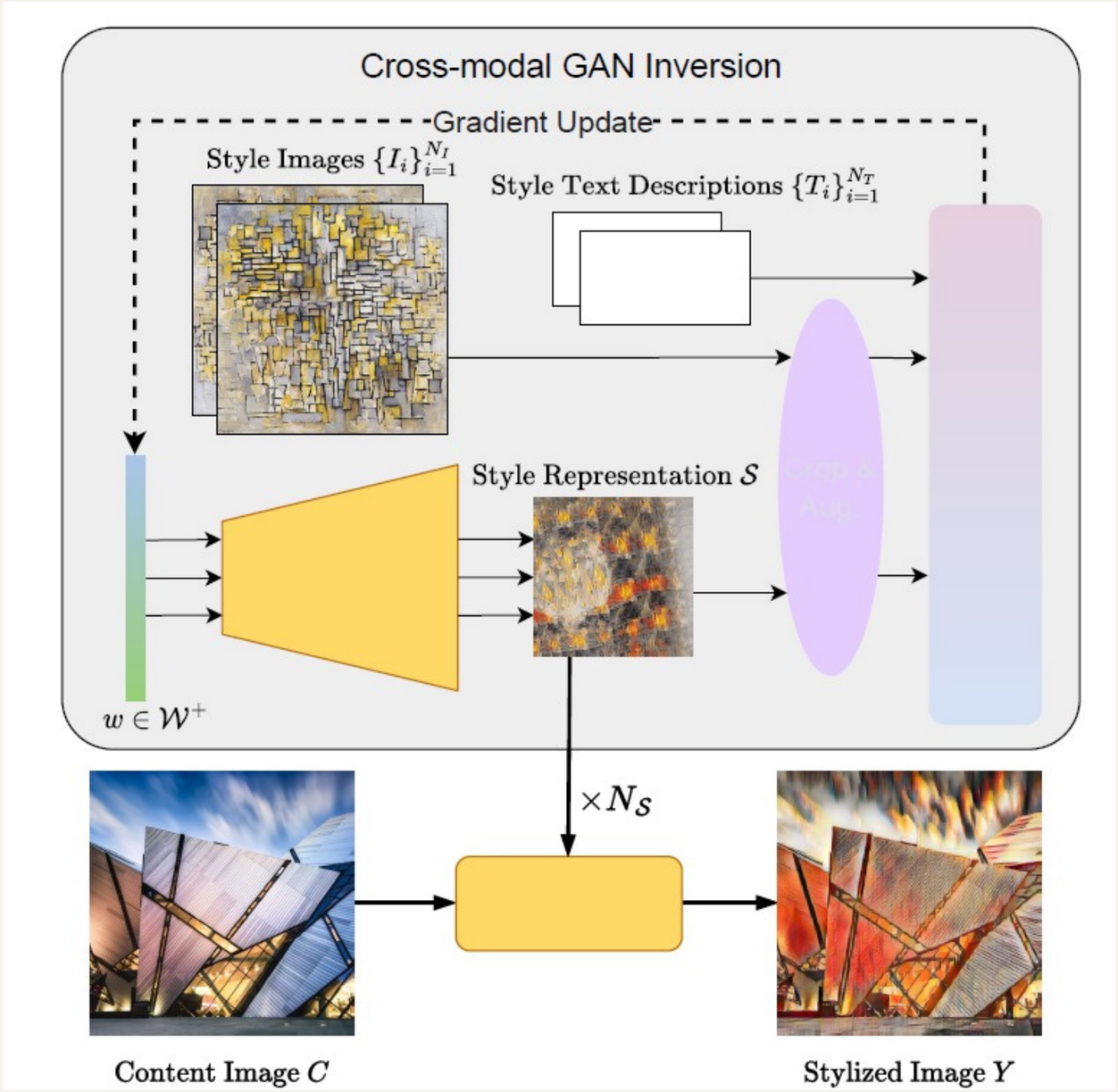
PROBLEM ADDRESSED

First Problem

CLIP-STYLER fails to disentangle the style and content information from both text and image.

Second Problem

Problem in describing the style by only text descriptions or only style image(Limited Style Guidance)


Cross-modal GAN Inversion
Gradient Update
Style Images $\{I_i\}_{i=1}^{N_I}$
Style Text Descriptions $\{T_i\}_{i=1}^{N_T}$
Style Representation $\mathcal{S}$
$w \in \mathcal{W}^+$
$\times N_S$
Content Image C
Stylized Image Y

METHODS USED

I.CROSS-MODAL GAN INVERSION:

It aims to combine different styles together to generate intermediate style representations. Therefore, the inversion should be able to accept multiple inputs from different references and modalities. Besides, only the style components of inputs are required to be inverted as their content parts are irrelevant to the downstream task.

I.I STYLE SPECIFIC CLIP LOSS

This paper wants the model to handle style inputs from the image modality as well. To this end,It propose an image-image patch-wise directional CLIP loss as below:

$$
\begin{aligned}
&\mathcal{S} = \mathbf{G}(w) \\
&\left\{\mathcal{S}^j\right\} j = 1^{N\text{crop}} = \operatorname{aug}(\operatorname{crop}(\mathcal{S})) \\
&\Delta \mathcal{S}^j = E_I\left(\mathcal{S}^j\right) - E_I\left(I_{\text{src}}\right) \\
&\Delta T = E_T\left(T_i\right) - E_T\left(T_{\text{src}}\right) \\
&L_{T_i} = \frac{1}{N_{\text{crop}}} \sum_{j=1}^{N_{\text{crop}}}\left(1 - \frac{\Delta \mathcal{S}^j \cdot \Delta T}{\|\Delta \mathcal{S}^j\|\|\Delta T\|}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\left\{I_i^k\right\} k = 1^{N\text{cop}} = \operatorname{aug}\left(\operatorname{crop}\left(I_i\right)\right) \\
&\Delta I_i^k = E_I\left(I_i^k\right) - E_I\left(I_{\text{scc}}\right) \\
&L_{I_i} = \frac{1}{N_{\text{crop}}^2} \sum_{j=1}^{N_{\text{crop}}} \sum_{k=1}^{N_{\text{crop}}}\left(1 - \frac{\Delta \mathcal{S}^j \cdot \Delta I_i^k}{\left\|\Delta \mathcal{S}^j\right\| \quad \Delta I_i^k}\right)
\end{aligned}
$$

In the general case where multiple style images {Ii}NIi=1 and style text descriptions {Ti}NTi=1 are given, we calculate the style-specific CLIP loss Lsty, and solve the following optimization problem

$$w^* = \arg\min_{w \in \mathcal{W}^+} L_{\text{sty}} = \arg\min_{w \in \mathcal{W}^+} \sum_{i=1}^{N_I} \alpha_i^I L_{I_i} + \sum_{i=1}^{N_T} \alpha_i^T L_{T_i},$$

I.2 INVERSION ALGORITHM

- Inversion algorithms start with an image and attempt to "invert" the process to recover a representation in the latent space of the generative model. This latent space captures the underlying factors that contribute to the style of the image.
- When both {Ii}NI i=I and {Ti}NTi=I are provided, crossmodal style interpolation, e.g., interpolating a style between given text and a given image, can be naturally achieved by adjusting the style weights {αIi }NIi=I and {αTi }NTi=I.

2.MULTIMODALITY-GUIDED IMAGE STYLE TRANSFER

- ## 2.1 Multi-Style Boosting:
  - This research proposes a method to improve the quality of style transfer in images. The issue addressed is that a single style representation might not capture all the nuances of the desired style.
  - The solution involves running a style transfer algorithm (cross-modal GAN inversion) multiple times for each set of style references. This creates a collection of various style representations. These representations are then used together to enrich the final result, while still being compatible with the adapted IIST model .

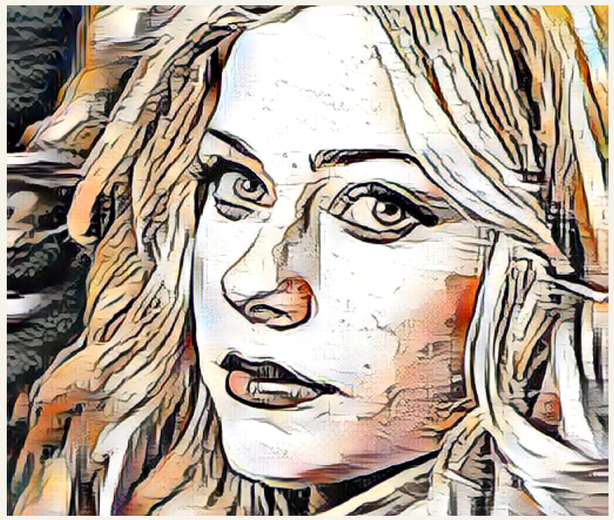
2.2 STYLE TRANSFER ALGORITHM

Regular image style transfer can miss stylistic details. This research tackles this by creating multiple variations of the desired style using a special algorithm. These variations are then used together to achieve a richer and more accurate style transfer in the final image. This method is also efficient, especially for processing images with similar styles

REPRODUCED RESULTS OF PAPER

---------->

TEXT INPUT:
CARTOON

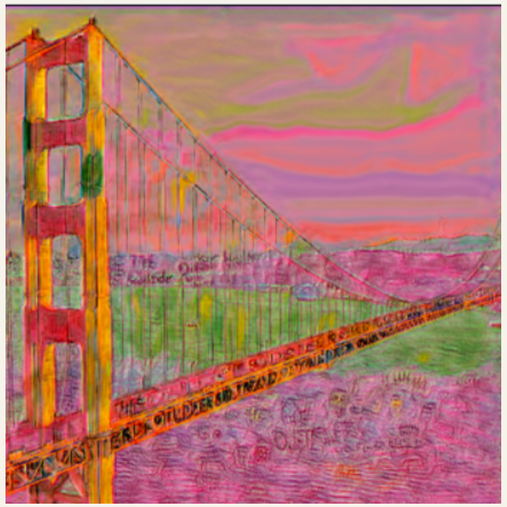


---------->

TEXT INPUT:
FIRE

QUALITY ANALYSIS:

	Brad Pitt (style fire)			Golden Gate (style Outsider Art)		
	CLIP	Spectral - CLIP	MMIST	CLIP	Spectral - CLIP	MMIST
SSIM	0.567	0.476	0.606	0.405	0.559	0.642



CLIP

Spectral CLIP

MMIST

USER STUDY:

For this we generated 3 different stylized mages(content+text pair) for each model.We asked participants(30) to rate each on a scale of 1-5.

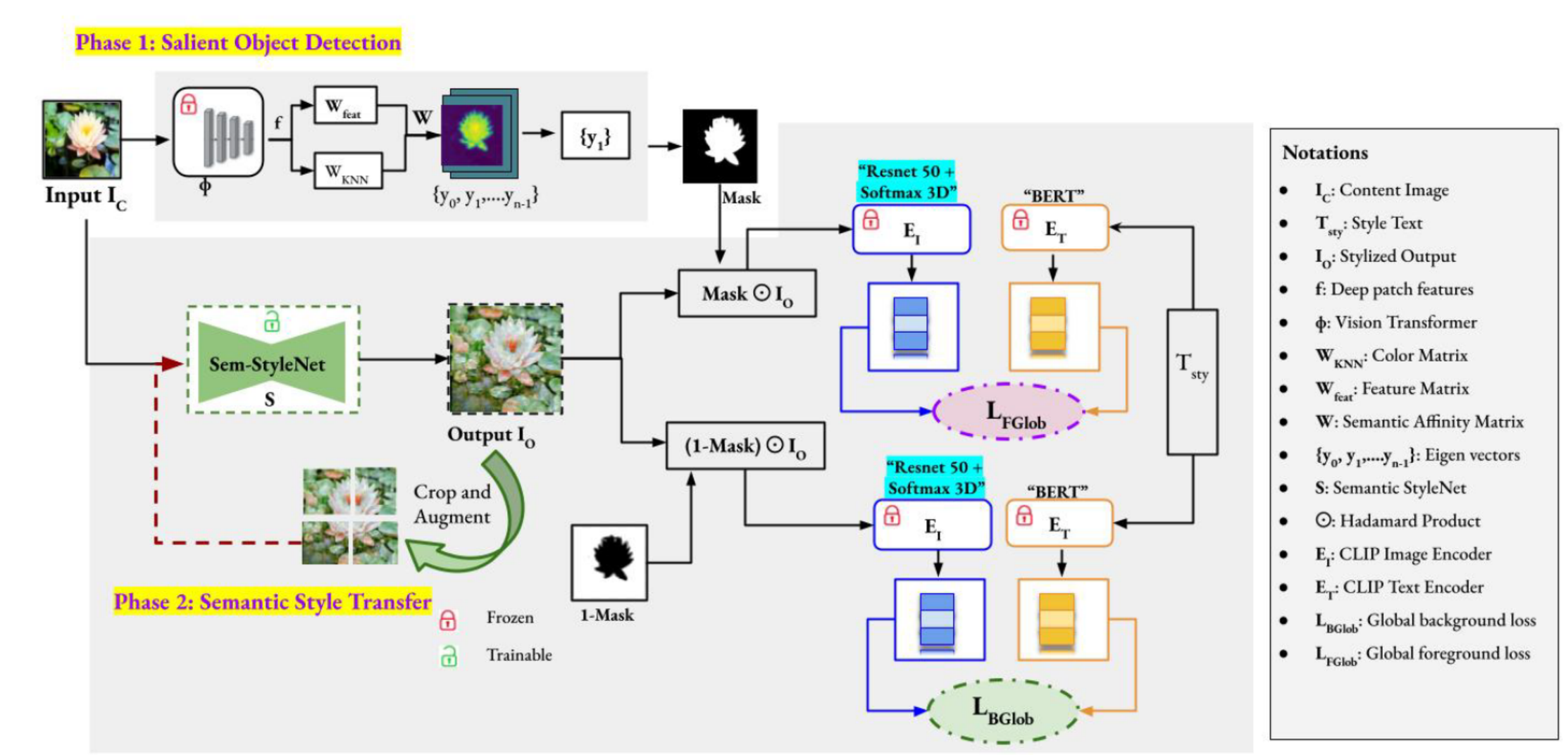
MODEL	IMAGE FIDELITY
CLIPStyler	2.90
SpectralCLIP	2.77
MMIST	3.83

Suggested Improvements

Idea 1) - Implementing style on Foreground and Background-

- When the MMIST model is implemented only using Text inputs, the resultant image lacks distinction between foreground and background. The edges are not smooth. Hence we solve this issue by implementing style individually and then combining the images to give us the resultant image with better edges.
- To perform this we follow the Sem-CS paper's directions and make some changes in it.
- So first we generate a mask and apply it to the image to generate 2 different images, ie Foreground, and background and we will treat them as 2 individual images.
- Now we will apply the process of Cross Modal GAN inversion on these 2 images.
- Finally, we will combine the result generated to give us the final Generated image.
- Doing this will help to separate background and foreground and we can even implement different styles on it, thus allowing us to stylize the image with more customization further.

Sem-CS Architecture



Notations

- I_C: Content Image
- T_{sty}: Style Text
- I_O: Stylized Output
- f: Deep patch features
- ϕ: Vision Transformer
- W_{KNN}: Color Matrix
- W_{feat}: Feature Matrix
- W: Semantic Affinity Matrix
- $\{y_0, y_1, \ldots y_{n-1}\}$: Eigen vectors
- S: Semantic StyleNet
- ⊙: Hadamard Product
- E_I: CLIP Image Encoder
- E_T: CLIP Text Encoder
- L_{BGlob}: Global background loss
- L_{FGlob}: Global foreground loss

PROJECT GANTT CHART

S. No	Task	Date of Completion
1	Understanding the Problem Statement	01/20/2024
2	[if any]:- Dataset Collection/ Pre-processing/ Literature review	01/30/2024
3	Baseline methods identification	02/08/2024
4	Working on project implementation	02/29/2024
5	Improving the accuracy by Hyper-parameters search	03/30/2024
6	Final Project Presentation Making	04/01/2024
7	Final Project Demo Making	

THANK YOU