DIB-TIST: Dynamic Image Blending for Enhancement of Text-based Image Style Transfer

Elishben Baraiya[1], Varad Bane[1], Anushka Jain[1], Harsha Rani[1], Indra Deep Mastan[2][0000−0001−5033−9561], and Preety Singh[1][0000−−0002−6885−1376]

[1] The LNM Institute of Information Technology, Jaipur 302031, India
{21ucs077,21ucc111,21ucc022,21ucs088, preety}@lnmiit.ac.in
http://lnmiit.ac.in
[2] IIT (BHU), Varanasi, India
indra.cse@itbhu.ac.in
http://https://iitbhu.ac.in/

Abstract. Text-based image style transfer (TIST) enables users to define styles through textual descriptions. A notable limitation in current TIST method of Multimodality-Based Image Style Transfer (MMIST) framework is the lack of distinction between the foreground and background in stylized outputs. To address this challenge, we propose a dynamic blending of the content images with the stylized image using a tunable hyperparameter. This is designed to compare regions with similar semantic meaning while considering the context of the entire image. We develop an optimal approach to integrate the appropriate proportion of content image features into stylized images, ensuring refined edges and enhanced perceptual quality. Our proposed approach yielded an Structural Similarity Index Measure (SSIM) of 0.606 and Gram matrix Difference (GMD) of 0.394. Our work offers an effective solution for improving the clarity and visual appeal of stylized images, contributing to the ongoing advancements in style transfer technologies.

Keywords: Style transfer · Text-based image style transfer · Contextual loss.

1 Introduction

Style transfer is a transformative technique in computer vision which enables the modification of an image's appearance by applying artistic characteristics of one image (referred to as *style image*) to another image (called the *content image*) [1]. It preserves the structure of the content image while seamlessly integrating stylistic features to it. It has numerous applications in real world like creation of digital art, enhancing marketing visuals to boost brand engagement and enriching the entertainment experiences by applying artistic styles to graphics and scenes.

Traditional style transfer techniques have proven effective but are often constrained by the requirement of a suitable reference style image. Text-based image

style transfer (TIST) extends the capabilities of style transfer by allowing users to define styles through natural language descriptions, such as *a watercolor effect* or *an abstract cubist painting* [4]. By leveraging the expressive power of language, TIST removes the dependency on reference images, providing unparalleled flexibility and accessibility. This innovation not only expands creative possibilities, but also normalizes style transfer technology, making it accessible to users without technical expertise. The collaborative use of deep learning and computer vision enable the system to process both visual and textual inputs effectively and achieve high-quality stylistic transformations.

TIST leverages vision-language models to generate stylized images based on text descriptions. These models encode both textual and visual representations into a common space, facilitating the association of texts with relevant images. This offers greater flexibility than Image-based Image Style Transfer (IIST). Vision-language models like CLIP (Contrastive Language–Image Pretraining) play a pivotal role in TIST [7]. By maximizing the similarity between aligned image-text pairs and minimizing it for mismatched pairs during training, CLIP learns a shared latent space for textual and visual representations. Many other methods, like CLIPStyler [4], were also developed to enhance the results of TIST, mainly with the aim of eliminating its limitations of having visual and textual artifacts due to entanglement in the embedding space of the vision language models.

Multimodality-Based Image Style Transfer (MMIST) framework leverages style guidance from multiple modalities [9]. It builds upon traditional IIST and TIST to synthesize stylized images that maintain the integrity of the content while achieving the desired artistic style. This method is particularly suitable when styles are not easily represented through a single reference. MMIST effectively addresses the limitations of TIST by reducing the undesirable visual and textual artifacts in the stylized images. It also offers a greater flexibility in style transfer by incorporating multiple style inputs from diverse modalities.

1.1 Motivation

Existing TIST techniques mainly use the large vision-language models like CLIP [7][4]. However, directly utilizing CLIP for style guidance often results in undesirable artifacts, such as the inclusion of unrelated visual entities or text fragments in generated images. This is mainly due to the image-text entanglement in the latent space of these vision-language models.

Many models have tried to address this issue [9][10]. MMIST allows style inputs from multiple sources and modalities and generating multiple style representations using GAN inversion technique. This not only resolves the issue of artifacts but also provides the added advantage of accommodating diverse stylistic inputs. However, if only the TIST component of MMIST is utilized, the stylized images generated often lack proper foreground and background separation. This limitation arises due to the suppression or blurring of edges from the content image, which leads to a loss of structural perception in the final stylized outputs.

To address these issues, we aim to achieve improved structural preservation in the stylized images, ensuring a more coherent separation between the foreground and background while maintaining the creative essence of the text-specified style. In the proposed work we seek to resolve this by allowing more appropriate structural integrity in stylized outputs as well as preserving the intended artistic style. Our main contributions are:

- Propose a novel blending of the style image output with the content image to enhance the stylized image.
- Conduct extensive experiments to determine the optimal blending weight by dynamically computing the contextual loss.

Our paper is organized as follows: Section 2 presents literature relevant to TIST and MMIST. Section 3 describes our proposed methodology ahile the experiments and results are outlined in Section 4. Section 5 concludes the paper.

2 Literature Review

In this section, we present few research work in the domain of TIST and MMIST. Text-Based Image Style Transfer enables users to define styles using textual descriptions instead of reference images, offering unparalleled flexibility [5],[2], [12]. CLIPStyler [4] introduced the first framework to achieve style transfer without a style image, relying solely on textual descriptions. It utilized the CLIP model, which maps text and images into a shared embedding space, enabling style representations to be extracted from text. CLIPStyler employed patch-wise directional loss to align textual descriptions with image content, achieving global and fine-grained style details.

Semantic CLIPStyler (SEM-CS) [3] addressed issues like style spillover and content mismatch by segmenting the content image into salient and non-salient objects. It applied styles selectively based on text descriptions to enhance semantic coherence. FastCLIPStyler [8] introduced optimization-free methods for faster, lightweight text-based style transfer, achieving efficient stylization in a single forward pass.

MMIST [9], builds upon both IIST and TIST methodologies. It offers several innovations that address the limitations of prior methods. Cross-Modal Style Representation encodes style inputs from multiple modalities (e.g., text, images) into a shared latent space, ensuring meaningful style and content separation. Adapted Style Transfer Model uses an adapted IIST framework to synthesize stylized outputs while preserving structural details of the content image. However, it also provides the flexibility to choose any IIST model for style transfer once the style representations are achieved. Multi-Style Boosting Mechanism uses the idea of multiple style representations of the same style enhancing the quality of style representations, ensuring accurate and detailed stylistic transformations.

We selected MMIST as our baseline due to its ability to achieve the disentanglement of text and image embeddings in the shared space allowing integration

of text and image inputs seamlessly, addressing artifact-related issues, and enabling flexible and high-fidelity style transfer. It provided a strong foundation for addressing some of the limitations of TIST and devising innovative solutions for future work.

2.1 Research Gaps

Current techniques in the output from TIST module of MMIST and other TIST approaches exhibit blurred edges, which makes it hard to separate the foreground from the background. This blurring reduces the sharpness of structural details especially at the boundaries. As a result, features from the content image get lost. The edges should keep their structural integrity from the content image. Instead, they often become too stylized and lose their distinct look. We illustrate this issue in Figure 1 with magnified images highlighting the blurred edge areas. This drawback shows we need a method to maintain the fine balance between content and style for each pixel. Such a method should preserve structural details while adding artistic style.

Fig. 1: A content image (left) stylized using *fire* text description. The output (right) from the TIST module of MMIST shows blurred edges (inset) in the stylized image.

3 Proposed Methodology

We propose a method to improve stylization outcomes by introducing a method for blending appropriate proportion of content and style. This proportion is decided by calculating the contextual loss between the original and stylized images. Contextual loss gauges how similar two images look, with a focus on keeping structural details intact [6]. Contextual loss measures the distance between the

corresponding features in the latent space that aims to ignore the spatial positions of the features, thus being able to tackle even the non-aligned data. The contextual loss will be captured by the high-level features (e.g. edges and structures) of the content and the stylized image obtained from the TIST model. A higher contextual loss will signify that the image has moved further from the original, which means that the output has retained fewer structural details. So, to maintain these structural details, the amount of content in the stylized image needs to increase.

We propose the concept of *blend weight* which dynamically adjusts the contribution of content and style for each pixel based on contextual loss. Our proposed approach is shown in Figure 2. For our content image, x, we generate a stylized image y using the textual description of style in the TIST module of MMIST. (This output will be referred to as the MMIST output in further discussions.) A pre-trained deep convolutional network is employed to extract feature representations from x and y. We focus on the deeper layers of the network as they capture high-level structural details essential for perceptual similarity. To ensure that y retains the structure of x, we compute the contextual loss, L_S, which measures perceptual similarity by identifying corresponding feature pairs between the two images. Unlike traditional pixel-wise losses, contextual loss is robust to non-aligned content, preserving structural integrity while allowing flexibility in style adaptation. A higher value of L_S, indicating greater dissimilarity, increases the emphasis on content preservation.

Using the computed contextual loss, we derive a blend weight, W_B, which dynamically adjusts the contribution of content and style in the final output. The blend weight, W_B, can be controlled by a tunable hyperparameter λ, ensuring a controlled balance between style and content. W_B is applied to refine the MMIST output, producing the enhanced stylized image, z, a weighted combination of the content image and the stylized image obtained from MMIST. This adaptive adjustment enhances structural coherence, improves perceptual consistency, and mitigates distortions, resulting in a more visually appealing and well-balanced style transfer.

3.1 Compute Contextual Loss:

The feature maps of x and y are obtained from deeper layers of a convolutional network. Firstly, the contextual similarity, $S(x, y)$, between two feature maps is computed. The concept of contextual similarity is based on the idea that two images are similar if their corresponding sets of features are similar (refer Figure 3) [6].

Contextual similarity, $S(x, y)$, between two feature maps is given as:

$$S(x, y) = \sum_{l} C(X^l, Y^l) \tag{1}$$

$$= \sum_{l} \frac{1}{N^l} \sum_{j} \max_{i} \mathrm{C}_{ij}^{l} \tag{2}$$

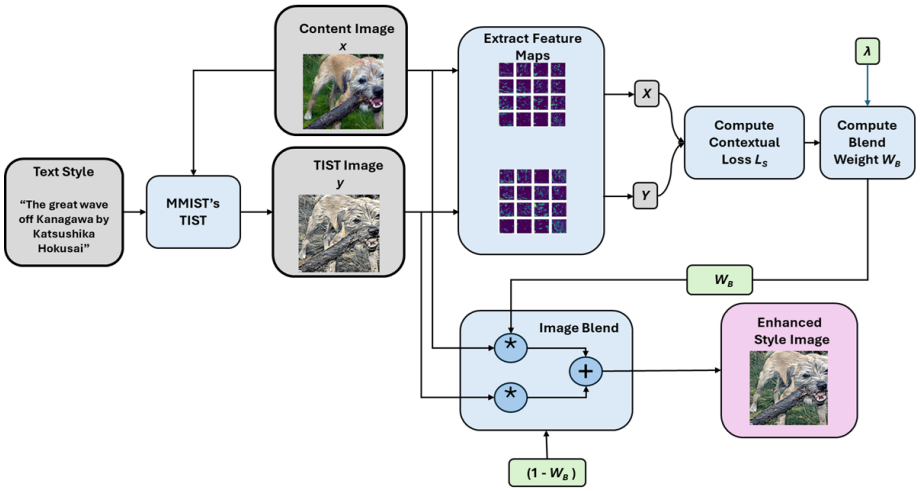


Fig. 2: Proposed methodology of DIB-TIST. The stylized image y obtained from the TIST module of MMIST is enahnced using a dynamlically adjusted blending of the content image x and y The blending weight depends upon the contextual loss and is controlled by hyperparameter λ.

where, $X^l(Y^l)$ are the set of feature maps of $x(y)$ extracted from layer l of the feature extraction network and N^l is the number of feature maps present in the l^{th} layer. C_{ij}^l is the contextual similarity between the i^{th} feature map of x, x_i^l, and j^{th} feature map of y, y_j^l, from the l^{th} layer. Here, $\max_i \mathrm{C}_{ij}^l$ ensures that for each feature map y_j^l, we find the feature map x_i^l that is most similar to it. Summing over all y_j^l gives the total contextual similarity for that layer. This is then averaged by dividing by the number of feature maps of that layer.

The contextual similarity score (C_{ij}^l) is calculated on the basis of the sum of the cosine distances between each pair of features obtained from the content image and MMIST output, of the same layer l of the network. The similarity between feature maps, x_i^l and y_j^l, is computed as:

$$d_{ij}^l = \left(1 - \frac{(x_i^l - \mu_y^l) \cdot (y_j^l - \mu_y^l)}{\|x_i^l - \mu_y^l\|_2 \|y_j^l - \mu_y^l\|_2}\right) \tag{3}$$

where, μ_y^l represents the mean of the feature maps of y from the l^{th} layer and $\|.\|_2$ computes the Euclidean distance between the vectors. The distance is normalized as follows:

$$\tilde{d}_{ij}^l = \frac{d_{ij}^l}{\min_k d_{ik}^l + \epsilon} \tag{4}$$

where, ϵ is a small constant to avoid division by zero and $\min_k d_{ik}^l$ represents the minimum cosine distance between a feature map x_i^l and any feature map of

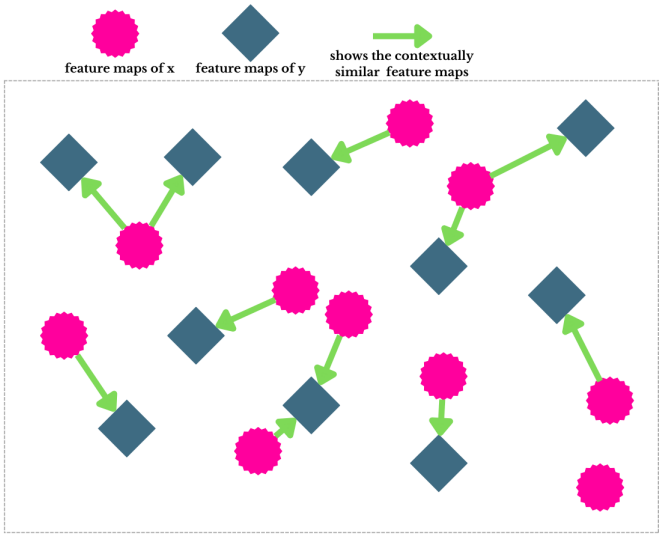


Fig. 3: Green diamonds represent the features of stylized image y while the pink circles represent the features of content image x. The arrows match each feature in y with its most contextually similar feature in x.

y at layer l. The normalized distance is converted to similarity:

$$w_{ij}^{l} = \exp\left(\frac{1 - \tilde{d}_{ij}^{l}}{h}\right) \tag{5}$$

where $h > 0$ is a bandwidth parameter that controls the sharpness of the similarity decay. A higher value of h results in a gradual decay of similarity whereas a lower value results in a sharper decay. For this implementation, $h = 0.5$ as recommended in the original paper [6].

The final similarity score between x_i^l and y_j^l is obtained by normalizing it to ensure that it lies between 0 and 1:

$$C_{ij}^{l} = \frac{w_{ij}^{l}}{\sum_k w_{ik}^{l}} \tag{6}$$

A higher C_{ij}^{l} indicates greater similarity between x_i^l and y_j^l. Contextual loss, L_S, is now computed as the negative log of the contextual similarity, $S(x, y)$).

$$L_S = -\log S(x, y) \tag{7}$$

3.2 Image Blending using *Blend Weight*:

As the obtained contextual loss, L_S, determines the amount of content to be added, the *blend weight*, W_B is derived from it as:

$$W_B = \lambda * clamp(L_S, 0, 1) \tag{8}$$

L_S is clamped between 0 and 1 to ensure that when multiplied with image pixels, it does not lead to values out of range. L_S captures the amount of mismatch in features between x and y. It is indicative of the loss in content and can help determine the amount of content image to be blended with the MMIST output. The hyperparameter, λ, gives control to maintain the proportion of content and style in the output. Using W_B, the enhanced stylized image, z, is computed as a weighted combination of x and y:

$$z = (1 - W_B) * y + (W_B) * x \tag{9}$$

If the contextual loss increases (indicating more difference between x and y), the blend weight also increases, pulling the target closer to the content. While this achieves our goal, it may lead to some style loss. Thus, λ, must be carefully tuned to balance content and style.

The quality of the enhanced image can be evaluated using two metrics: Structural Similarity Index Measure (SSIM) and the Gram matrix Difference (GMD) [1]. SSIM reflects any major degradation in the structural information which emphasizes about strongly inter-dependent pixels. It is computed between the enhanced image z and the original reference image x. SSIM can have a maximum value of 1, exhibiting maximum content similarity. Maximizing the content translates to $(1 - SSIM)$ being minimized.

GMD is used to represent the style of an image. Style refers to the overall texture and pattern distributions of the image, independent of their exact position. It is based on the idea that CNNs extract hierarchical features, each layer detecting some patterns like edges, texture or shapes. The style is a relationship between these patterns, and is determined by capturing the correlations between different features of the CNN layers. A difference in the gram matrix of the generated image with that of the style image will yield the style loss. We want this difference to be minimized to ensure good style transfer. In our experiments, we compute the Gram Matrix Difference score between the enhanced stylized image, z, and the intermediate internal style representations of MMIST. The focus of our approach is to get an enhanced image with low GMD.

4 Experiments and Results

Our experiments focused to enhance the content features of the TIST image obtained from MMIST. This is done by dynamically blending the content image and MMIST output based on the contextual loss.

4.1 Dataset

The Flickr30k dataset [11] is a popular benchmark for sentence-based picture portrayal comprising of 31,783 images. We chose a random subset of 500 content

images from the dataset. Few sample images are shown in Figure 4. Each image has a descriptive caption. It depicts a wide range of real-world activities, people, and objects in different contexts, allowing us to test the generalization of our method to different content structures when applying styles.

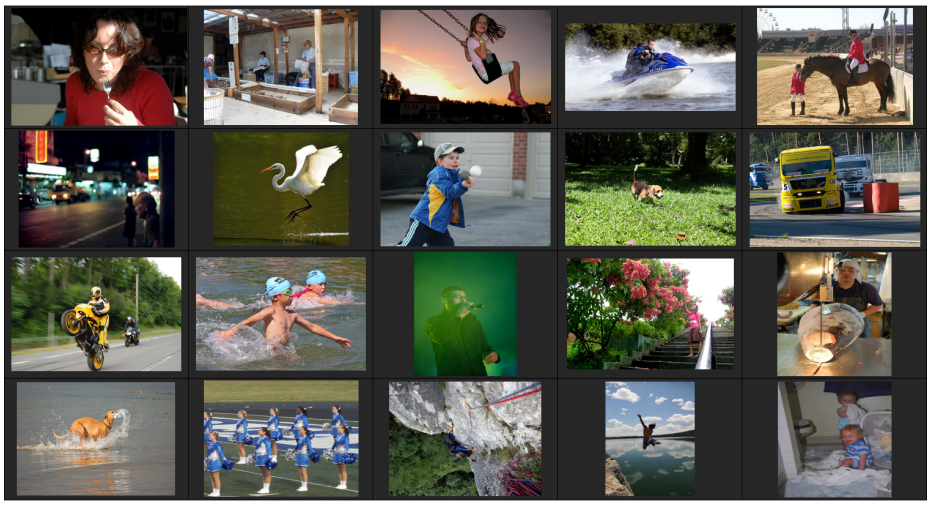


Fig. 4: Few sample images from the Flickr30k dataset [11]. Images show various real-world activities.

4.2 Extraction of Feature Maps

For the dataset images, stylized images were extracted from MMIST using five text styles, viz., viz. *fire*, *pop-art*, *mosaic*, *the great wave off Kanagawa*, and *white wool*. For the dataset images (content images) as well as the MMIST output images, feature sets were obtained from $Conv4\_2$ and $Conv5\_2$ layers of a pre-trained VGG19. These provide feature maps of sizes $28 \times 28 \times 512$ and $14 \times 14 \times 512$ respectively.

4.3 Tuning of Hyperparameter λ

We experimented with five different text styles. The content images from the dataset were combined with MMIST images to generate enhanced stylized images for each experimental value of λ, viz. 0.2, 0.4, 0.6 and 0.8 using Equation 9 which is influenced by λ as shown in Equation 8. Thus, 500 enhanced stylized images were generated per style form for each value of λ. We then computed the SSIM and Difference of Gram Matrix scores between the MMIST outputs and our enhanced style images. The results shown in Table 1 provide an average

value over the enhanced style images. A value of $\lambda = 0$ will yield the original stylized image from MMIST since this makes the blend weight, $W_B = 0$ and thus, zero content will be added. The qualitative results are shown in Figure 5.

Table 1: Table displaying different λ values with corresponding SSIM and Gram Matrix differences (GMD) for various styles.

	Proposed Approach			**MMIST Output (Averaged over Images)**	
Style	λ	**SSIM**	**GMD**	**SSIM**	**GMD**
Fire	0.2	0.572	0.271	0.313	0.137
	0.4	0.710	0.292		
	0.6	0.804	0.367		
	0.8	0.931	0.456		
Pop-art	0.2	0.391	0.095	0.244	0.068
	0.4	0.581	0.148		
	0.6	0.722	0.271		
	0.8	0.951	0.342		
Mosaic	0.2	0.511	0.289	0.562	0.164
	0.4	0.683	0.302		
	0.6	0.732	0.368		
	0.8	0.921	0.401		
The great wave	0.2	0.463	0.412	0.226	0.351
off Kanagawa	0.4	0.520	0.591		
	0.6	0.612	0.615		
	0.8	0.725	0.741		
White Wool	0.2	0.449	0.249	0.166	0.290
	0.4	0.538	0.439		
	0.6	0.679	0.678		
	0.8	0.919	0.919		

The graphs for $(1 - SSIM)$ and GMD scores for various values of λ are shown in Figure 6. As can be observed, the best value of λ corresponds to the intersection point of $(1 - SSIM)$ and GMD scores. SSIM increases as more content proportion is taken (corresponding to higher value of λ). However, this increase leads to an increase in GMD as well since the style depreciates. Hence, the best λ can be taken as 0.4 as it maintains a balance between these two metrics. However, this value is specific to this dataset and may require tuning for other images.

4.4 Comparison with Baseline MMIST Approach

Using our best value of $\lambda = 0.4$, we compared our results with the baseline approach, which is the output obtained from the TIST module of MMIST. We averaged the scores over the five styles for 500 content images. Our proposed approach yielded an SSIM of 0.606 and GMD of 0.394 while values obtained for

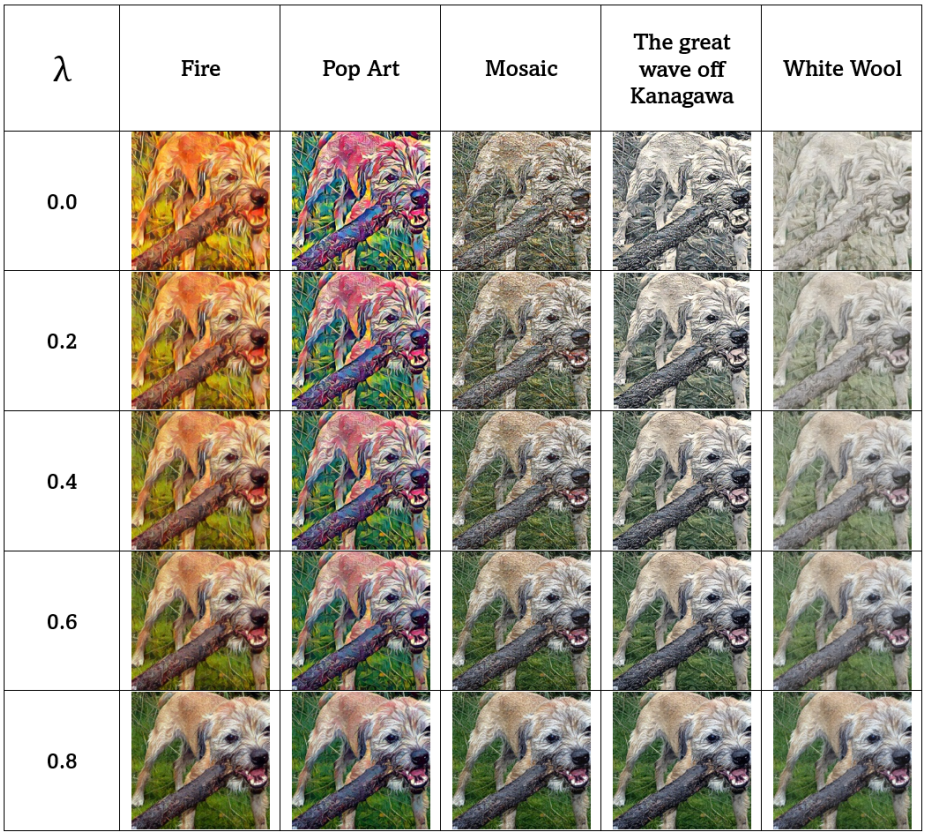
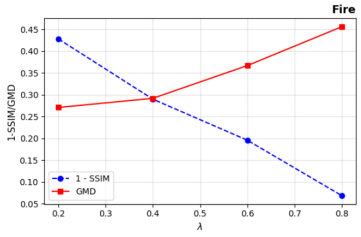


Fig. 5: Qualitative analysis of enhanced stylized image outputs based on tuning of hyperparameter λ for various styles.

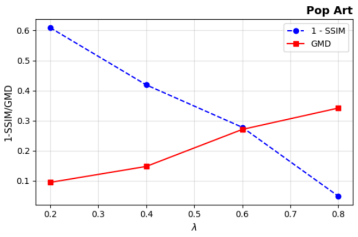
baseline images are 0.302 and 0.202 respectively, clearly showing the efficacy of our method to balance the content and style. It enhances the style of the image while maintaining good structural similarity with the content image.

5 Conclusions

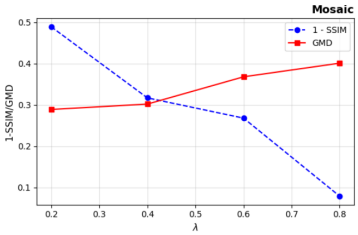
In this paper, we present a method to enhance the results of existing style transfer frameworks like the TIST module of MMIST by addressing its limitations. Our method focuses on adjusting the balance between the content and the style of each pixel using a blend weight from contextual loss. This improves the foreground-background separation and preserves the important structural details while also ensuring artistic style fidelity. Through detailed experimentation and extensive hyperparameter tuning, we propose a value of $\lambda = 0.4$ for a balanced content and style output. However, this value is specific to this dataset and may



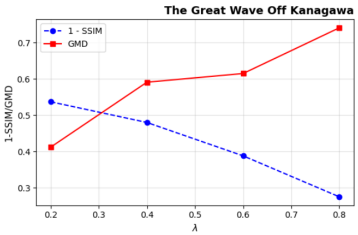
(a) Fire



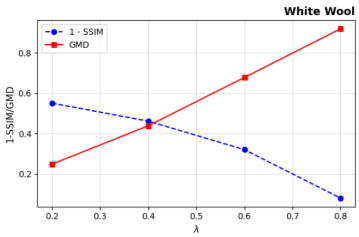
(b) Pop-art



(c) Mosaic



(d) The great wave off Kanagawa



(e) White wool

Fig. 6: Graph of $(1 - SSIM)$ / GMD scores for various values of λ for various styles. The goal is to minimize both these values. For most of the styles, the intersection is achieved at $\lambda = 0.4$ which yields a balance of both content and style.

require adaptive tuning for other images. Our approach improved stylization quality, particularly in edge regions. While this approach improved results by adding structural details, it is possible that more advanced deep learning-based methods can allow for even more sophisticated and meaningful blending. This may further enhance the integration of content and style into a seamless fusion. Future work may be undertaken on larger and diverse datasets with more styles and adaptive hyperparameter tuning.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2414–2423 (2016)
2. Huang, N., Zhang, Y., Tang, F., Ma, C., Huang, H., Dong, W., Xu, C.: Diffstyler: Controllable dual diffusion for text-driven image stylization. IEEE Transactions on Neural Networks and Learning Systems (2024)
3. Kamra, C.G., Mastan, I.D., Gupta, D.: Sem-cs: Semantic clipstyler for text-based image style transfer. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 395–399. IEEE (2023)
4. Kwon, G., Ye, J.C.: Clipstyler: Image style transfer with a single text condition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18062–18071 (2022)
5. Liu, Z.S., Wang, L.W., Siu, W.C., Kalogeiton, V.: Name your style: Text-guided artistic style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3530–3534 (2023)
6. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 768–783 (2018)
7. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
8. Suresh, A.P., Jain, S., Noinongyao, P., Ganguly, A., Watchareeruetai, U., Samacoits, A.: Fastclipstyler: Optimisation-free text-based image style transfer using style representations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7316–7325 (2024)
9. Wang, H., Wu, P., Rosa, K.D., Wang, C., Shrivastava, A.: Multimodality-guided image style transfer using cross-modal gan inversion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4976–4985 (2024)
10. Xu, Z., Xing, S., Sangineto, E., Sebe, N.: Spectralclip: Preventing artifacts in text-guided style transfer from a spectral perspective. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5121–5130 (2024)
11. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)

12. Zhang, Y., Huang, N., Tang, F., Huang, H., Ma, C., Dong, W., Xu, C.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10156 (2023)