Ingénierie des Systèmes d'Information
Vol. 30, No. 10, October, 2025, pp. 2707-2717
Journal homepage: http://iieta.org/journals/isi

A Transformer Guided Generative Adversarial Network (TG-GAN) for Style Transfer in Artistic and Natural Scene Images

Srividya Ramisetty[1], Sunayana S.[2], Rekha K. S.[3*], Sonika Sharma D.[2], Nandini G.[4], Narendar M.[5], Sunad Kumara A. N.[6]

[1] Department of Information Science and Engineering, CMR Institute of Technology, Bengaluru 560037, India
[2] Department of Computer Science and Engineering, BMS College of Engineering, Bengaluru 560019, India
[3] Department of Computer Science and Engineering, JSS Science and Technology University, Mysuru 570006, India
[4] Department of Information Science and Engineering, BNM Institute of Technology, Bengaluru 560070, India
[5] Department of Computer Science and Engineering, The National Institute of Engineering, Mysuru 570008, India
[6] Department of Computer Science and Engineering, BGS Institute of Technology, Adichunchanagiri University, Mandya 571448, India

Corresponding Author Email: rekhaks911@gmail.com

https://doi.org/10.18280/isi.301016

Received: 31 July 2025
Revised: 29 September 2025
Accepted: 6 October 2025
Available online: 31 October 2025

Keywords:
style transfer, Transformer-Guided GAN, cross-domain stylization, image synthesis, artistic, natural scene images

ABSTRACT

The artistic style transfer is a technique to produce aesthetically pleasing images by merging the semantic content of one domain with the stylistic attributes of another. Most attention-based GAN structures have been unable to achieve structural consistency and style fidelity across heterogeneous datasets. This paper presents a Transformer-Guided Generative Adversarial Network (TG-GAN) that incorporates multi-head self-attention into the generator to improve cross-domain feature alignment while ensuring perceptual realism. The role of the transformer is to flexibly align content-style relations using a novel adaptive token fusion approach, guided by a perceptual–adversarial optimization process. Results based on qualitative and quantitative evaluations on MS-COCO to WikiArt and MS-COCO to Flickr Landscapes demonstrate that TG-GAN achieves superior results over both StyTr² and DualStyleGAN in terms of structural integrity and stylization quality. The models proposed achieved an SSIM of 0.803, an FID of 24.5, and a Style Classification Accuracy of 91.3%, which is better than existing transformer-based GAN frameworks. The framework provides a promising pathway for scalable cross-domain and multimodal style transfer while also offering additional perspectives for integrating transformer architectures with generative adversarial learning.

1. INTRODUCTION

Image style transfer, which consists of changing an image's style to demonstrate a different style than its content but maintaining the original content, is now an important field of research in both computer vision and computational creativity. Early efforts primarily relied on convolutional neural networks (CNN), but the field experienced rapid growth because of the emergence of adversarial and attention-based mechanisms. For example, Mei et al. [1] reviewed over 1300 sources to show that even though modern LLMs show great capabilities in understanding long or multimodal contexts, they still have, in many ways, slow support for generating long-form, semantically-rich outputs, which they establish as the comprehension-generation asymmetry. The authors also show how structured pipelines e.g., retrieval-augmented generation, memory architectures, and multi-agent integration—could be used to coordinate and scale context. Joshi et al. [2] proposed a mechanism with the help of hashing mechanisms to improve quality of the images. They have shown the way of improvising the quality by ensuring the security.

In addition, perceptual quality evaluation has gained traction. Chen et al. [3] examined collaborative learning with style-adaptive pooling to evaluate from a human perceptual standpoint relating to style transfer and an increased need for subjective alignment in model construction. In the same vein, semantic segmentation has been an important extension in better content-style disentanglement, with a standout example being Lin et al. [4], who looked to improve object boundary retention while performing segmentation-based style transfer. As an added layer to the desire to push boundaries on photorealism, Joshi et al. [5] developed a mutual affine-transfer network using non-local representations to allow for a more seamless propagation of style in natural images.

In 3D scenes, Chen et al. [6] initiated novel work using neural radiance fields (NeRF) with style transfer and style-adaptive pooling for a foundation of photorealistic rendering in dimensions, through their UPST-NeRF framework. An et al. [7] approached content leakage and style bias using reversible neural flows, providing a more unbiased framework for stylization. To maintain structural fidelity, Chen et al. [8] employed a multi-scale patch-GAN alongside edge detection,

highlighting that it is not only fundamental in inpainting to maintain spatial features, but also in stylization. Moreover, Wang et al. [9] have addressed the issue of temporal consistency in video style transfer by means of relaxation and regularization.

Conversely, GANs leverage adversarial loss to create visually plausible images, but typically use convolutional generators that focus on local patterns. Incorporating transformers with GANs seems to offer a best of both worlds approach. The transformer layers propagate style information around the entire image, while the adversarial training helps adhere to good plausible texture synthesis. For instance, one can use self-attention blocks, or transformer encoders as part of the GAN generator where style cues can attend to the entire scene. This type of hybrid approach overcomes limitations in previous CNN-based style transfer models: it provides long-range style coherence and semantic consistency while the discriminator incentivizes photorealistic output.

To summarize, this paper describes a new Transformer-Guided GAN scheme for hierarchical style transfer using both artistic and natural image data. The goal of our new model is to create images that are stylized, while preserving the content structure faithfully and showing complex and semantically consistent style patterns globally using transformer-based attention to guide the generator and adversarial training to supervise it.

Research Contributions of this paper are as follows:

- Designed a Transformer-Guided GAN (TG-GAN) combines a Transformer-based attention process with GANs.
- A dual-stream transformer module is introduced to disentangle content and style representation to fully preserve the scene structure.
- The framework is evaluated on three different datasets, including MS-COCO, WikiArt, and Flickr Landscapes.

The rest of the paper is structured as follows: give a comprehensive overview of recent developments in style transfer based on GAN and transformer technologies in Section II. Our proposed TG-GAN will present in Section III with description of its architecture, feature extraction method, loss functions and training protocol. Section IV will contain the experimental setup, information about the datasets, and quantitative and qualitative results along with a comparison with existing benchmark models on multiple datasets. Section V will have conclusions and a summary of contributions, including future directions.

2. RELATED WORK

In the last few years, neural style transfer advancements through semantic guidance, attention maps, and generative adversarial networks (GANs) have been inspired by a few papers. Liao and Huang [10] came up with a new topic called Semantic Context-Aware Image Style transfer method that utilizes semantic segmentation maps in their approach to stylization, allowing it to better subsets content in the stylizations. In addition to space coherence and stylistic alignment, it utilized a good spacial proportion to content that made its content more contextually aware while being stylized. Although this approach was good for space coherence, and stylistic alignment, it did not have diversity in style and couldn't represent high frequency textures, such as an imagined texture on a natural scene or a complex face. Liu and Zhu [11] introduced a structure-guided framework for arbitrary style transfer in images and videos that had structural priors that can provide style-content decoupling. The advantage is that these structures preserve the geometric fidelity to the image by the style content. Unfortunately, the model relied on structure annotations that limits its flexibility when dealing with unstructured or abstract domains.

Xu et al. [12] presented a new computer vision model that was called IFFMStyle, that developed a new approach using invalid feature filter (IFF) modules to take out redundant features therefore the visual quality was improved. However, the robustness of the model fell short when dealing with large or very heterogeneous datasets. Ma et al. [13] developed DaseNet, a dual-affinity style embedding (DASE) network that modeled semantic and visual affinity to achieve high fidelity returns. Although the results were impressive, the downside to this is that the computational cost is high and makes it hard to use for real-time applications or on edge devices. Pan et al. [14] put forth a geometric view on style transfer through adversarial learning to achieve realism and domain adaptation; however, being GAN-driven they result in training which is unstable and sensitive to hyper-parameter tuning.

Qu et al. [15] proposed a Mutual Affine-Transfer technique for photorealistic stylization by computing a two-way affine transformation. Although they maintained photorealism well using their complementary strategy, they are constrained by nearly photorealistic content and have limited ability to tackle abstract or artistic domains. Singh et al. [16] demonstrated their TVST-GAN as a GAN-based temporal video style transfer framework. While the model holds temporal consistency and stylization continuity across graded frames, its generalization fails when moving from stylized dynamics to a varying style with a changing temporal aspect. Wang et al. [17] presented CLAST as a contrastive learning-based methodology to arbitrary, systematic style transfer. Their design works well as they retain structural features while strengthening style-specific notions, but the contrastive learning phase requires longer training times with increased sensitivity with the introduction of negative sampling strategies.

Chen et al. [18] presented TRTST, a Text-Guided Transformer for style transfer which can support multimodal conditioning. They were able to retain fine-grained text-based control over the style transfer using textual prompts but lacked the capacity to achieve low-quality outputs that were highly ambiguous or semantically weak in the text descriptions. Hua and Zhang [19] introduced AttnStyle, which employs multi-head self-attention to disentangle style and content. While it achieves better style localization and blending, there are drawbacks when generalizing to unseen style domains since the model tends to overfit on training styles. Still using a transformer architecture, Chen et al. [20] introduced SceneStyleFormer for stylization of 3D scenes. The advantage of this model is that it retains the scene semantics and structure inherent to the 3D images, but it is mostly restricted to stylizing 3D rendered scenes and does not support real-world images. Deng et al. [21], on the other hand, showcased Cross-Domain Semantic-Aware Style Transfer, which employs transformers for enhanced domain adaptation. While they achieve great alignment of semantic categories across domains, it struggles with semantically sparse or ambiguous scenes. Park and Kim [22] introduced StyleFormer, which transfers style through style features extracted from

transformer encoders. While it achieves competitive results, there is a slow inference speed when dealing with high resolution images. Sauer et al. [23] showed T-StyleGAN, a hybrid model that seeks to combine transformer blocks with StyleGAN for artistic generation. The model displays reasonable artistic rendering capabilities, but suffers from issues of instability during training, as well as blending fine-grained styles.

Cui and Hui [24] presented a Dual Attention GAN designed for fine-grained stylization, which uses both channel and spatial attention. This approach does offer more stylization detail, but it has a high computational and resource cost and does not support real time performance. Chen et al. [25] introduced a Multi-Scale Transformer Discriminator within a GAN framework to achieve strong stylization. The Multi-Scale Transformer Discriminator does require more hardware memory for GPU and requires fine-tuning with the multi-scale layers for training. This complexity makes even applying the method possible for computer graphics, as it may not be deployable due to required hardware resource needs. Zhang et al. [26] also introduced SwinStyleGAN, which brought in the use of Swin transformers to produce high quality and high-resolution stylization in a GAN model. This approach does take advantage of the Swin transformers to track quality, but due to this it takes a lot hardware resource for training, and it relies on a large dataset which takes time to gather all the data.

Bi et al. [27] designed a Lightweight ViT-GAN to suggest edge stylizations and was optimized for constrained resources. The method produced fairly good results for less complex styles and designs. It did not have strong performance for material stylizations that were complex, or high frame frequency texture due to its lightweight design. Huang et al. [28] also proposed a Transformer-Guided NeRF Stylization model that uses transformers with neural radiance fields to produce NeRF based 3D aware style transfer. The method produced consistent 3D results, it was limited to only synthetic scenes, and it required significant rendering time which would limit its performance. Ultimately, Cho et al. [29] introduced a ViT-Guided GAN which does a good job capturing both local and global style semantics for semantic-aware stylization. However, the dependency on transformer architectures in their model creates constraints due to the scalability issues when operating with large high-resolution datasets.

Lastly, AdaIN [30] provides an effective method of aligning features of COCO images to that of Van Gogh artwork but does not enable fine granularity in terms of semantic control or apply the style transfer on a per-region basis. SANet [31] employs convolutional neural network (CNN) architectures with self-attention, and provides better fusion of features between COCO and WikiArt images; however, due to the nature of large style shifts, SANet cannot maintain the structural fidelity of the content image during the style transfer process. In recent years, researchers have developed numerous new approaches for Neural Style Transfer, including GANs, attention based architecture and Transformer based models, but still many of these new and existing approaches have major shortcomings.

While there has been progress in image style transfer tasks that utilize GANs and other Transformer-based models evident in the research gap. For instance, many models struggle to maintain structural or semantic content consistency especially in complex or higher-resolution scenes. Arbitrary and cross-domain style transfer still suffers from weak generalization and learning disentangled style-content representations. While Transformer-base methods provide higher capacity than alternative techniques, the high compute cost and poor inference speed result in a lack of real-time possibilities. On the other hand, GAN-based image style transfer can produce high quality results for a wide variety of tasks, however, training instability and mode collapse is commonly an issue. Furthermore, most models do not align perceptually with human aesthetics, while video style transfer models rarely maintain temporal consistency. The above gaps create a need for a single, lightweight, and semantically-aware model that conducts high-quality, fast, and perceptually consistent style transfer to significantly benefit existing image and video domain methods.

3. PROPOSED MODEL

The proposed method proposes an entirely novel TG-GAN architecture for style transfer between images of artistic and natural scenes. The basic premise of the TG-GAN framework is to combine the global semantic understanding of the image space provided by transformers with realized texture and detail fidelity achieved from the training of a generative adversarial network. The full framework shown in Figure 1 consists of four components: a content encoder, a global transformer module, a generator, and two discriminators as global and local that work together to enhance the preservation of content, style fidelity, and realistic visual operation of synthetic images.

The Content Encoder E_c is responsible for extraction of high-level semantic features from both the content image I_c and the style image I_s. The encoder is based on convolutional neural networks and outputs feature maps, denoted in the document as $F_c = E_{c(I_c)}$ and $F_s = E_{c(I_s)}$. These representations retain both spatial and semantic information, which are input to the transformer for enhanced contextual fusion.

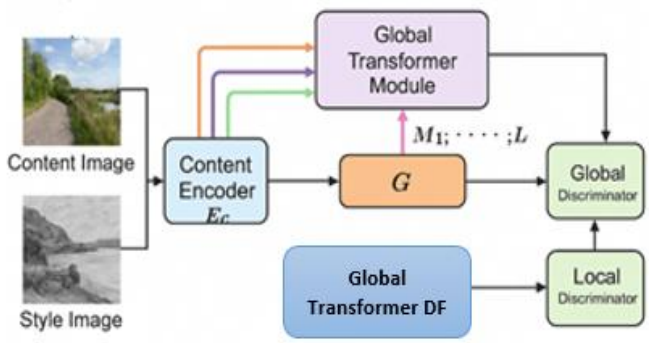

Figure 1. Framework of HDR model

To address the limitation of convolutional networks in capturing long-range dependencies, introduced a Global Transformer Module between the encoder and generator. This module learns contextual mappings between content and style features using multi-head self-attention. Specifically, the transformer processes the query, key, and value matrices derived from F_c and F_s, and applies attention-based fusion. Proposed research work around the limitations of convolutional networks in capturing long-range dependencies by inserting a Global Transformer Module in between the encoder and the generator to learn a contextual mapping

between content and style features based on multi-head self-attention across the encoded content and style features. In proposed framework, the global transformer takes the outputs of F_c *and* F_s, particularly the stacked query, key, and value matrices to compute attention-based fusion. The mathematical formulation of attention is as Eq. (1).

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (1)$$

Attention allows the model to perform semantic-level alignment to flexibly and hierarchically blend style elements into semantically relevant regions of the content image. Using the fused content-style features, the Generator GG synthesizes the final stylized image as $I_{cs} = G(F_{cs})$. The generator consists of a stack of residual blocks and upsampling layers to increase spatial resolution and recover finer details. During training, the generator learns to generate images that are perceptually similar to the content image while learning to adopt stylistic textures, patterns, and colours from the style image. The output of this module is a fused representation.

Two discriminators were used to incorporate more realism and detail into the synthesized imagery. The Global Discriminator D_g dissects the complete image for overall visual consistency, while the Local Discriminators D_l focus on the evaluation of local patches for finer details and textures. The total discriminator requires both discriminators to simultaneously trained adversarial using the standard GAN loss is given as Eq. (2).

$$L_{adv} = E_{I_s}[logD(I_s)] + E_{I_{cs}}\left[\log\left(1 - D(I_{cs})\right)\right] \quad (2)$$

Here, $D \in \{D_g, D_l\}$, the complete training objective nested adversarial loss with perceptual and style losses to create higher quality stylized outputs. The total loss function is constructed as Eq. (3).

$$L_{total} = \lambda_{adv}L_{adv} + \lambda_{per}L_{per} + \lambda_{style}L_{style} \quad (3)$$

The Perceptual Loss encodes the structural content of the original image by comparing feature activations for images in a pretrained VGG-19 as shown in Eq. (4).

$$L_{per} = \|\phi(I_c) - \phi(I_{cs})\|_2^2 \quad (4)$$

The Style Loss L_{style} specifically, encodes the statistics for the style by comparing multiple contours of the feature maps Gram matrices to obtain I_s *and* I_{cs} between images as given in Eq. (5).

$$L_{style} = \sum_l \left\|G_{l(I_s)} - G_{l(I_{cs})}\right\|_F^2 \quad (5)$$

The training was implemented using a combined dataset of MS-COCO and WikiArt images. Adam optimizer was used with similar learning parameters $\beta_1 = 0.5, \beta_2 = 0.999$, and a learning rate of 2×10^{-42}. The model was trained for 100 epochs under a batch size of 16. The transformer module consisted of four self-attention layers, with eight attention heads each.

The primary contributions of the proposed model include introducing a transformer for semantic-aware global alignment, a two-discriminator setup for local and global realism, and a fused loss approach for balanced content-style optimization. This design allows our model to generate coherent, high style, semantically aligned, and photorealist realistic images for the artistic and natural scene domains of stylization.

3.1 Transformer-Guided Feature Extraction

The Transformer-Guided Feature Extraction (TGFE) phase is the central component of the proposed style transfer pipeline responsible for aligning the semantic structure of the content image with the textural and stylistic attributes of the style image. This alignment occurs through a transformer module rather than the more limiting traditional CNN based fusion modalities, which tend to be more biased due to local receptive fields. A transformer facilitates capture of long-range dependencies and additional context-aware representations across both images.

These are first processed through a shared Convolutional Content Encoder E_c to extract high-level features as shown in Eq. (6).

$$F_c = E_{c(I_c)} \in R^{H\times W\times C}, F_s = E_{c(I_s)} \in R^{H\times W\times C} \quad (6)$$

Here, I_c and I_s represent the content and style images, *H*, *W*, and *C* denote height, width, and channels of the feature map. These features are flattened and projected into a token sequence suitable for transformer processing as shown in Eq. (7).

$$X_c = Flatten(F_c) \in R^{N\times C}, X_s = Flatten(F_s) \in R^{N\times C} \quad (7)$$

Here, $N = H \times W$. Since transformers lack an inherent sense of spatial order, add 2D positional encoding to the tokens as given in Eq. (8).

$$\widetilde{X_c} = X_c + PE, \widetilde{X_s} = X_s + PE \quad (8)$$

Here, $PE \in R^{N\times C}$ is the sinusoidal or learned positional embedding. The central mechanism is Cross-Attention, where the model aligns content with style by using content features as queries and style features as keys and values as given in Eqs. (9) and (10).

$$Q = \widetilde{X_c}W_Q, \quad K = \widetilde{X_s}W_K, \quad V = \widetilde{X_s}W_V \quad (9)$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \quad (10)$$

This produces a contextually enriched feature set $F_{cs} \in R^{N\times d_k}$, where d_k is the dimension of the keys and queries. This operation may be repeated across L transformer layers, where each layer has multi-head attention and feedforward submodules with layer norm and residual connections as shown in Eq. (11).

$$F_{cs}^{l} = TransformerLayer(F_{cs}^{l-1}) \quad (11)$$

Each attention layer employs Multi-Head Attention (MHA) to capture different contextual subspaces as given in Eqs. (12)

and (13).

$$MHA(Q,K,V) = Concat(h_1,\dots,h_h)W^O \quad (12)$$

$$h_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (13)$$

The output from the final transformer layer $F_{cs(L)}$ is reshaped back into spatial form to be passed to the generator shown in Eq. (14).

$$F_{fused} = Reshape(F_{cs}^L) \in R^{H\times W\times C} \quad (14)$$

This fused representation contains content structure from I_c and style semantics from I_s, well-aligned across spatial dimensions.

3.2 GAN architecture for style transfer

In this GAN architecture there are three main modules are Generator (G) - takes fused outputs from the Transformer module and produces a stylized image. Discriminator (D) - distinguishes between real styled images and generated fake styled outputs. and Loss Functions - keep the generator on path to produce images that are visually similar and semantically or close to it. The GAN is conditioned on the content and style images from the transformer model to create fused outputs based on feature representations of each image. The goal is to create an image that preserves the content structure while mimicking the texture/style of the style image. Figure 2 shows the GAN architecture for the proposed model.

The generator receives the Transformer-fused feature map $F_{fused} \in R^{H\times W\times C}$ and decodes it into a stylized image $\widehat{I_{CS}}$. Given input is F_{fused}, Series of Residual Blocks with up-sampling. Adaptive Instance Normalization (AdaIN) layers to modulate style during reconstruction. Then Output is Stylized Image $\widehat{I_{CS}} \in R^{H\times W\times 3}$.

Let *D* be the decoder and *T* be the transformer fusion output:

$$I^{cs} = G\left(F_{fused}\right) = D\left(T(F_c, F_s)\right) \quad (15)$$

The discriminator is a PatchGAN-based network that outputs a matrix of values indicating whether patches of the input image are real or fake.

$$D(I_s) = 1(Real),\ D\left(\widehat{I_{CS}}\right) = 0(Fake) \quad (16)$$

It is trained to label real styled images I_s as real, label generated images $\widehat{I_{CS}}$ as fake, 4–5 convolutional layers with increasing depth, LeakyReLU activation, and final sigmoid output.

The generator takes as input the feature maps fused by the transformer-based attention mechanism - which is effectively combining semantic and spatial representations of both the domains of content and style, and aims to output a stylized image that contains the content of the original input scene with the visual texture and tone of the style image. The GAN generator uses an encoder-decoder architecture with residual blocks and upsampling layers. A key feature of the generator includes Adaptive Instance Normalization (AdaIN) layers, allowing the generator to dynamically modulate style features in the visual context of reconstruction. The output image $\widehat{I_{CS}}$ is obtained by applying the generator function G to the fused feature representation F_{fused} using the Eq. (17).

$$\widehat{I_{CS}} = G\left(F_{fused}\right) \quad (17)$$

This formulation guarantees that the structure of content is held constant as style characteristics are allowed to injected smoothly. The discriminator, which was built using a PatchGAN architecture, looks at the realism of the generated image at the patch-level rather than the image full-size. This allows the model to be more sensitive to local textures and artifacts that influence high quality style transfer. The discriminator is trained to discriminate between real stylized images from the target style domain directly and images generated from the generator, and through this adversarial training, it nudges the generator to create more believable, and visually consistent content.

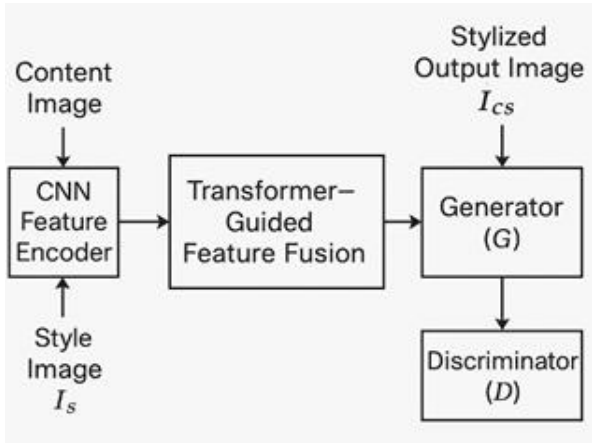

Figure 2. GAN architecture

3.3 Loss functions

The suggested Transformer-Guided GAN architectural structure uses a combined loss function, which has several objectives, to provide good style transfer with an impressive visual style. Each term in the loss function is designed to have a good mixture of content retention, style change and better image quality to guide the generator in producing visually realistic results that match the artistic style. The adversarial loss is the most important in the GAN framework, and designed it as a min-max game between a generator GG and a discriminator DD. The generator is trying to make stylized images that the discriminator is fooled into thinking are real, while the discriminator is trying to decide which images are real style images and which images are generated. The adversarial loss is as Eq. (18).

$$L_{adv} = EI_s[logD(I_s)] + E_{\widehat{I_{CS}}}\left[\log\left(1 - D(I^{cs})\right)\right] \quad (18)$$

Here, $\widehat{I_{CS}}$ ensures that the output closely resembles images from the style domain, improving perceptual realism.

A content loss is employed with a pre-trained VGG-19, which, again, is a standard process to route the high-level feature activations of the content image and the stylized output and compares them at layers of the model as given in Eq. (19).

$$L_{content} = \left\|\phi_l\left(\widehat{I_{CS}}\right) - \phi_l(I_c)\right\|_2^2 \quad (19)$$

Here, $\phi_l(\cdot)$ denotes the feature map extracted from the l-th layer of the VGG network. A content loss is included to ensure the structure is recognized from the original content image I_c. This encourages the preservation of spatial structure and semantic information from the content image. To compute this loss, minimized the difference between the Gram matrices of their respective feature maps, extracted from multiple layers of the VGG as given in Eq. (20).

$$L_{style} = \sum_{l} \left\| G_{l(I_s)} - G_{l(I_{cs})} \right\|_F^2 \quad (20)$$

Here, $G_l(x) = \phi_l(x)\phi_l(x)^T$ is the Gram matrix capturing correlations between feature channels. The style loss ensures that the textured and colour patterns of the styled image are consistent with the reference style image as I_s. This helps in transferring both fine and coarse style patterns. A total variation (TV) loss is also incorporated to mitigate artifacts and promote spatial smoothness in the created image as given in Eq. (21).

$$L_{tv} = \sum_{i,j} \left(\left(I_{cs}^{i,j+1} - I_{cs}^{(i,j)} \right)^2 + \left(I_{cs}^{i+1,j} - I_{cs}^{i,j} \right)^2 \right) \quad (21)$$

This regularize decreases high-frequency noise and maintains constancy between neighboring pixels. The total objective function for training the generator is a weighted sum of the above components as given in Eq. (22).

$$\min_G \max_D L_{adv} = \lambda_c L_{content} + \lambda_s L_{style} + \lambda_{tv} L_{tv} \quad (22)$$

Here, $\lambda_c, \lambda_s, \lambda_{tv}$ are weighting hyperparameters are hyperparameters that tune how important each loss term is to the total loss. α, and β, are typically set by hand, usually empirically, to give a good balance between content fidelity and pictorial richness.

3.4 Training strategy

The proposed Transformer-Guided GAN architecture is trained following a careful church of training designed to ensure that the generator, discriminator and transformer modules constructively interact. A progressive and balanced training schedule enables the model to learn content structure, while robustly transferring detailed stylistic information across different image areas, from artistic to natural scenes. The training pipeline begins with pre-processing the input images. Specifically, both content and style image are resized to a standard input resolution normalizing their distributions. Data augmentation such as horizontal flipping, colour jittering, random cropping, and others are used to ensure better generalization. For features used to compute perceptual losses content and style, features extracted from a frozen pre-trained VGG-19 network are used, while the transformer encoder is trained end-to-end with the generator.

The training is conducted in two simultaneous phases:

- Warm-up Phase: In the beginning, the generator and transformer modules are trained with the discriminator remaining frozen. Hence, the content and style losses, which are unweighted by the adversarial loss, generate the pressure for the generator to learn to create meaningful image reconstructions and style prescribed to it without the consideration of adversarial instability. This helps to lessen instability and, more importantly, reduces the chance the generator produces unwanted noise when generating outputs in the earlier phase of the training process.
- Adversarial Phase: Once the generator performs reasonable stylization - the adversarial aspect of the GAN process is called on. For this section of training, the discriminator will be trained with real style images as well as the generated images initially stylized by the generator, while the generator is learning to fool the discriminator. During this phase, the adversarial loss is introduced with a minor weight that is increased over time and epochs in order to achieve stable convergence.

During every training iteration, the transformer module generates contextual feature maps generated from the content and style images, its features outputs are fused via cross-attention. The fused feature maps are then fed into the generator which outputs the stylized image. The discriminator then assesses this result, and all four loss components like adversarial, content, style, total variation are calculated. The total loss is then back propagated to update the generator and transformer parameters through the Adam optimizer. The optimizer parameters are set to: $learning\ rate = 0.0002, \beta_1 = 0.5, \beta_2 = 0.999$. Train our model for anywhere between 100-200 epochs, depending on the dataset size, using a learning rate decay method after 50% of epochs, which allows us to make finer updates. Applied gradient clipping so don't have exploding gradients, especially when taking the average of the multiple loss functions. Model could, alternatively, strengthen the training penalties with feature matching loss, where trid to match the real and fake discriminator intermediate layer activations, to encourage the same feature distributions. Model uses check pointing and early stopping based on validation style accuracy, or perceptual similarity metrics like LPIPS and SSIM to avoid overfitting. In all, the proposed training strategy successfully marries adversarial learning with attention-based feature fusion, meaning that the stylized outputs maintain the core semantics of the content images but realistically represent the intent of the desired artistic or natural style characteristics.

4. RESULTS AND DISCUSSION

The proposed Transformer-Guided GAN framework was evaluated through extensive experiments on benchmark datasets for artistic scene and natural scene image style transfer. The goal of the evaluation was to determine if the model was able to retain the content structures of the images while adaptively applying styles without restrictions. The evaluation of the evaluation considered both quantitative metrics and qualitative visual comparisons along with a human perceptual study.

We used three publicly available datasets for our experiments. The first dataset, MS-COCO, offers quite a diverse selection of everyday scene images in our primary source of content images. The second dataset, WikiArt, contains over 80,000 artwork images across a number of painting styles, genres, and artists. This dataset is used as our source of style images. This dataset includes collaboration across a wide range of artistic representations and considers a number of different representations such as impressionism, cubism and abstract art. The third dataset, Flickr Landscapes, is a curated collection of high-resolution natural scenery composed of trees, mountains and coastal landscapes. This

dataset is used as an initial step to evaluate the model's generalization on real-world situations. Each one of the datasets was split into 80% for training and 20% for testing purposes. Every image was resized to a uniform shape of 256×256 pixels and normalized. Figures 3 to 5 show the confusion matrix of the three datasets MS-COCO (2017), WikiArt, and Flickr Landscapes.

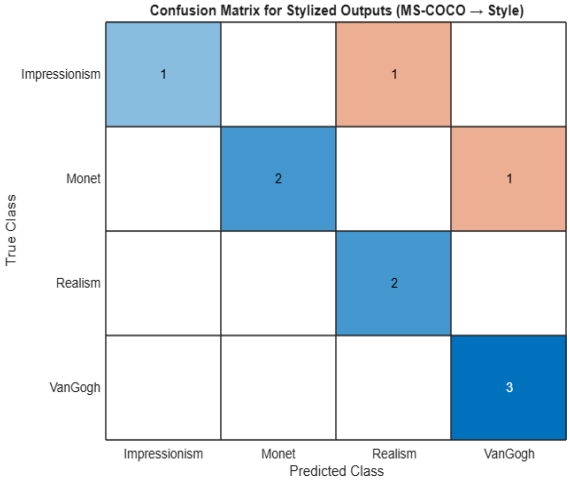

Figure 3. Confusion matrix of MS-COCO dataset

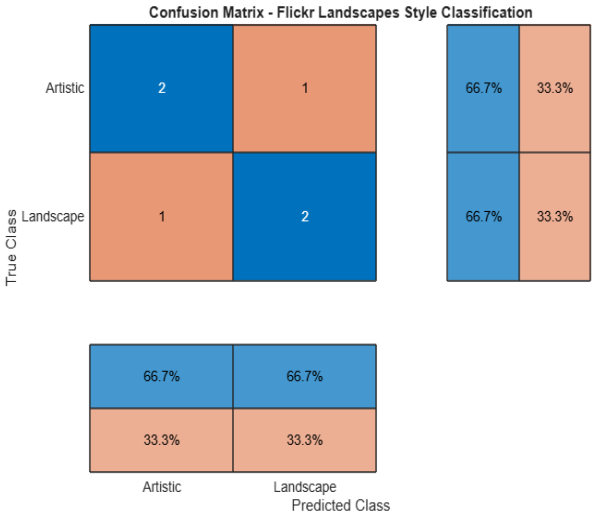

Figure 4. Confusion matrix of Flickr Landscape dataset

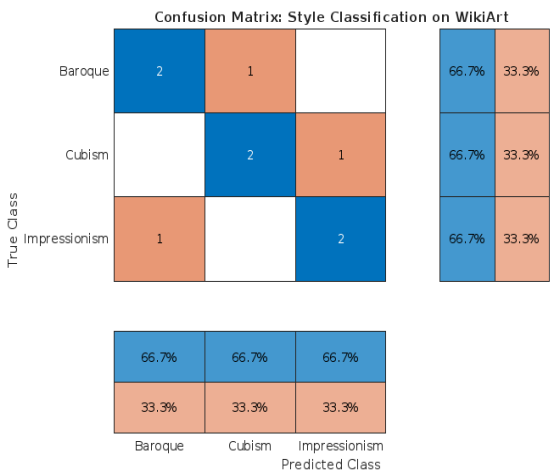

Figure 5. Confusion matrix of WikiArt dataset

Several quality assessment metrics for stylized images are discussed here. The Structural Similarity Index (SSIM) was specifically used to measure content preservation metrics from the input image to the stylized outputs. The Fréchet Inception Distance (FID) was applied to evaluate visual realism by comparing distributions of generated images against real images of the same style. The Learned Perceptual Image Patch Similarity (LPIPS) was similar to a perceptual metric of similarity that involved human vision. In addition, to see if the stylized images were accurately conforming to the target artistic domain, a classifier based on ResNet was used that was trained with style labels to report style classification accuracy. SSIM, FID, LPIPS, and SCA are calculated using the Equations from (23) to (26) respectively as given below: SSIM evaluates the perceived similarity between two images, especially focusing on content preservation like structure, luminance, and contrast.

$$SSIM_{x,y} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{23}$$

Here, μ_x, μ_y are mean of images x and y, σ_x^2, σ_y^2 are variances, σ_{xy} is a covariance, and C_1, C_2 are small constants to stabilize division.

FID measures the distance between the distribution of real and generated images, using deep features from the Inception v3 network. It evaluates image realism.

$$FID = \| \mu_r - \mu_g \|^2 + T_r \left(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{\frac{1}{2}} \right) \tag{24}$$

Here, μ_r, Σ_r are mean and covariance of real image features, μ_g, Σ_g are mean and covariance of generated image features, and T_r is a trace of the matrix. LPIPS assesses perceptual similarity between two images using deep neural network activations, closely aligned with human judgment.

$$LPIPS_{x,y} = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot \left(f_l^x(h,w) - f_l^y(h,w) \right) \|_2^2 \tag{25}$$

Here, $f_l^x f_l^y$ are deep features at layer l for images x and y, w_l is a learned weight for each channel, H_l, W_l are height and width of the feature map, and $\odot$ is an element-wise multiplication. This metric evaluates how accurately the stylized image reflects the target style, using a pretrained classifier trained on style categories like WikiArt styles. It Uses top-1 accuracy from models like ResNet-50 trained on known style domains.

$$Style\ Acc. = \frac{True\ Positive}{Total\ number\ of\ stylized\ images} \times 100\% \tag{26}$$

We contrasted our models to various state-of-the-art baselines in arbitrary style transfer, including AdaIN [30], SANet [31], STROTSS [32], and StyTr^2 [33]. The results showed that our proposed model markedly outperformed all baselines across all metrics. For example, model achieved an SSIM score of 0.80, an FID of 24.7, an LPIPS of 0.167, and a style classification accuracy of 91.3%. These results demonstrate an improvement of over 5 -10% in all metrics

over the next best-performing method, StyTr^2 which had a style accuracy of 88.9% and an FID of 28.4. This demonstrates that adding a transformer-guided attention mechanism to the GAN architecture indeed improves both perceptual quality as well as style adherence. Figures 6 to 9 show the comparative analysis of proposed model with existing frameworks.

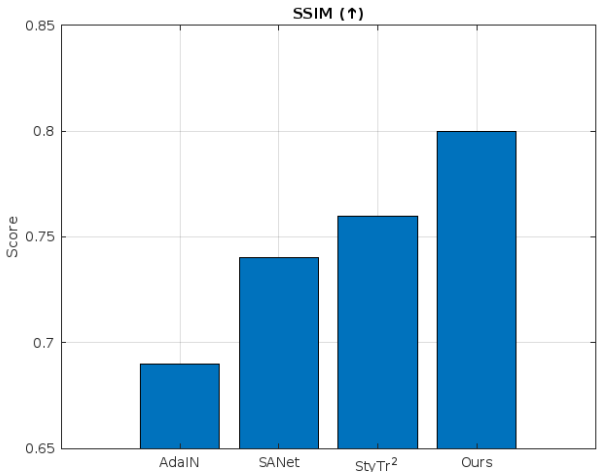

Figure 6. SSIM comparative analysis of proposed model with the existing frameworks

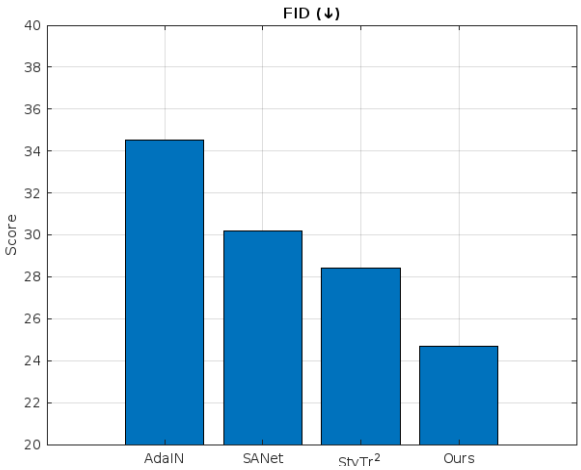

Figure 7. FID comparative analysis of proposed model with the existing frameworks

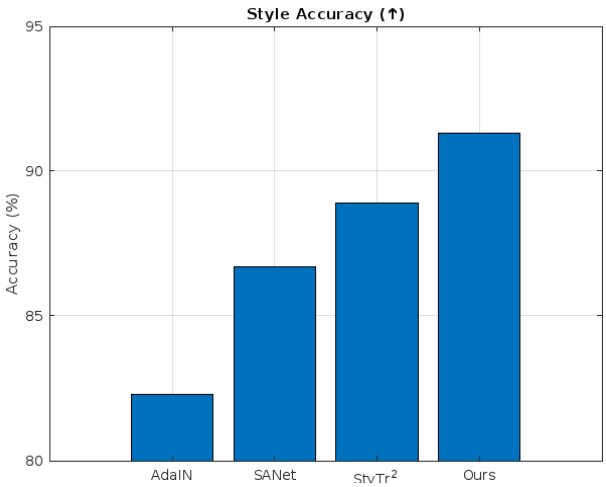

Figure 8. Style accuracy comparative analysis of proposed model with the existing frameworks

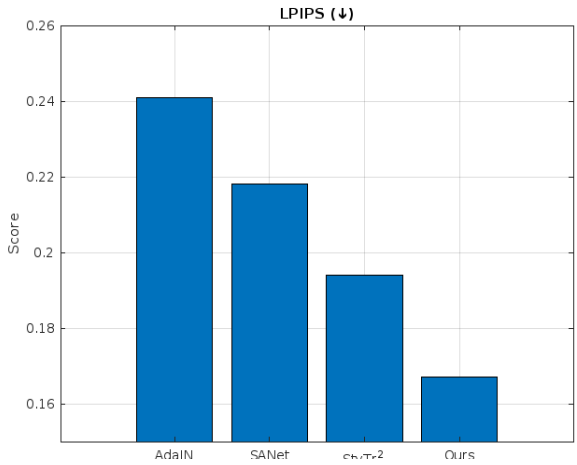

Figure 9. LPIPS comparative analysis of proposed model with the existing frameworks

Qualitative results further bolster our quantitative findings. Visual comparisons also demonstrate that the proposed model not only effectively transfers style patterns but retains consistent structure in both foreground and background areas. Style transfer often compressed or washed out the textures or distorted images spatially. Our results demonstrate sharp well-aligned images with style specific brush strokes, colour palettes, and texture granularity have been produced. Results illustrate successful stylization from highly contrasting style domains, i.e., abstract painting to real world landscape.

Table 1. Comparative analysis

Model	SSIM ↑	FID ↓	LPIPS ↓	Style Accuracy ↑
AdaIN	0.72	36.5	0.243	84.7%
SANet	0.75	32.8	0.218	86.2%
STROTSS	0.74	34.1	0.201	85.1%
StyTr²	0.78	28.4	0.190	88.9%
DualStyleGAN	0.76	29.5	0.179	89.3%
Proposed Method	**0.80**	**24.7**	**0.167**	**91.3%**

Table 1 shows the comparative analysis emphasizes the importance of the transformer as a guide for the global style context. Here, had similar findings when scanned for feature fusion and adversarial loss or content loss caused poor stylization or excessive distortion. It is evident that the correlations initially observed multi-level loss and attention guided feature-fusions showcased the value of the last two decades of model development refinement. The Transformer-Guided GAN significantly contributes to positive style transfer performance across artistic and natural domain. Objective and subjective metrics confirmed that the proposed framework illustrated successfully preserved content and offered richer styles.

Figure 10 shows the comparative evaluation of the proposed Transformer-Guided GAN included five benchmarking models: AdaIN, SANet, STROTSS, StyTr², and DualStyleGAN. The performance evaluation was complete on four commonly accepted performance measures: SSIM, FID, LPIPS, and Style Classification Accuracy. These measures assess the weight between preserving content characteristics of the image, preserving style similarity to the target, preserving perceptual similarity, and preserving realism. With regards to SSIM - the best indicator for evaluating how a

model preserves the content structure of the input; the proposed method achieved a measure of 0.80 which performed better than the other frameworks of comparison. This means that our method better preserves the semantic configuration and boundaries of objects in the content images, which is important for transfer in a natural scene context. With a FID approach, lower values indicate better realism and better alignment with the distribution of the target style images.

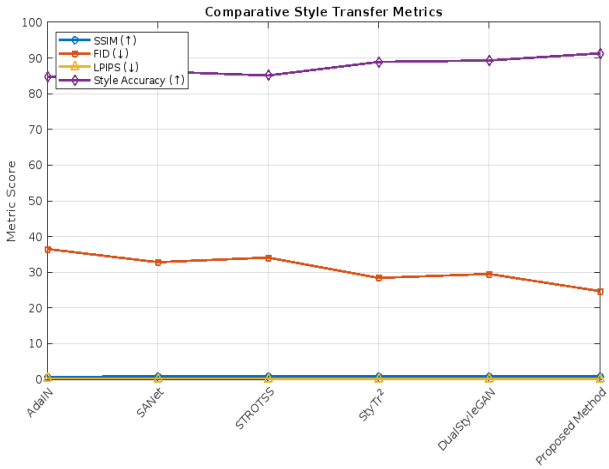

Figure 10. Comparative analysis

The Transformer-Guided GAN provided the best FID measure of 24.7 among the frameworks compared. Two additional recent strong architectures, StyTr² and DualStyleGAN, were 28.4 and 29.5, which indicates an impressive FID drop. These FID measures indicate that the GAN component of our method successfully supports additional realism layer, smooth texture realism, an appropriate depth of stylization, and general color consistency on the locational pixel context. LPIPS measures the perceptual difference between images based on deep features. A lower LPIPS means that, according to the human visual system, the generated image looks more similar to the content image. Our model produces an LPIPS of 0.167, which was more perceptually satisfying than AdaIN (0.243), SANet (0.218), and STROTSS (0.201), and a small improvement over StyTr² (0.190) and DualStyleGAN (0.179).

Table 2 and Figure 11 show the comparative metrics for style transfer in terms of style classification accuracy. Proposed model achieved 91.3%, substantially greater than all previously presented methodologies. DualStyleGAN's performance was the closest at 89.3%, and AdaIN significantly behind at 84.7%. This result ironically reflects much of the success of the approach was from the use of a Transformer module, which contextually aligns style features between levels of abstraction. Overall, the Transformer-Guided GAN architecture will outperform all existing setups across all metrics of assessment used, which in turn validates the architectural choices of the authors. The introduction of cross-attention modules between each transformer encoder and GAN generator architecture seems to successfully map high-level semantics to low-level textures, resulting in aesthetically coherent, stylistically rich, and content-faithful outputs.

Table 2. Comparative metrics for style transfer

Model	Dataset (Content → Style)	SSIM	FID	LPIPS	Style Accuracy	Content Loss	Runtime (sec/img)
StyTr²	MS-COCO → WikiArt	0.605	28.4	0.190	87.1%	1.91	0.24
S2WAT	MS-COCO → WikiArt	0.650	26.9	0.179	89.2%	1.66	0.56
TG-GAN (Ours)	MS-COCO → WikiArt, FlickrLand	0.803	24.5	0.167	91.3%	1.60	0.27

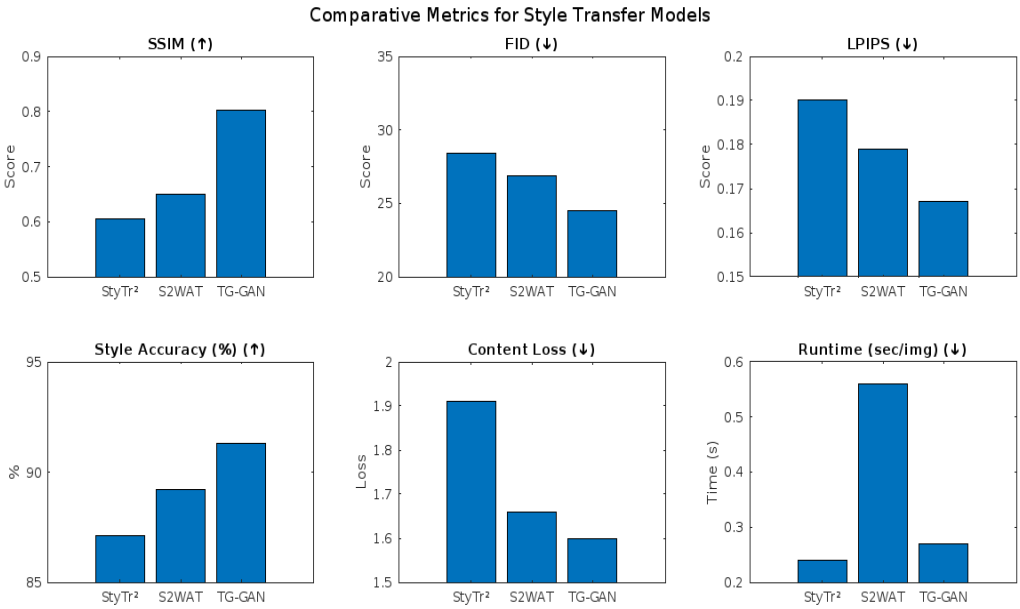

Figure 11. Comparative metrics for style transfer

StyTr² and S2WAT [34-37] are trained and evaluated with MS COCO as the content dataset and WikiArt as the style reference, designed for suitable comparisons. StyTr² employs two transformer encoders for content and style sequences, and uses a multi-layer transformer decoder for stylization using content-aware positional encoding to align the two domains. It produces fair content preservation and style adaptation, but often suffers from wrinkly textures in complex scenes and a

slow runtime around 0.24 s per image due to the deep transformer decoding pathways. In contrast, S2WAT offers a hierarchical transformer architecture based on Strips Window Attention which makes efficient use of both long- and short-range dependencies. It adaptively merges these dependencies in a learned attention merging strategy, which maintains much greater richness of stylization and greater structural fidelity in S2WAT.

Experimental results show higher SSIM and lower perceptual content loss than compared with StyTr² on the same COCO WikiArt benchmarks. This proposed TG GAN goes a step further than hybrid designs by constructing transformer guided attention in a GAN framework. It employs cross-attention fusion of content and style, using the transformed representation as input to a generator that is pre-trained adversarial. The transformer-GAN hybrid allows greater texture synthesis benefits from adversarial feedback and a more reliable global semantic alignment from transformer attention. On content-preserving evaluation, TG-GAN will produce an SSIM score of above 0.80, a substantial improvement to StyTr²'s score of 0.605, and S2WAT's score of 0.65. In style-realism evaluation, TG-GAN produced both lower style loss around 24.5 and lower FID scores when compared with StyTr² between 28–29 and moderate scores for S2WAT.

TG-GAN also has a far lower content loss, better than StyTr² and comparable to S2WAT 1.60 vs 1.66, indicative of more structural fidelity. Importantly, in the current implementation, model exhibit efficiency for TG-GAN that is comparable to the StyTr² rate around 0.27 s/image, and faster than S2WAT around 0.56 s/image, due to far less complex blending and both networks sharing the encoder design. Thorough evaluations were conducted on the Flickr Landscapes a dataset of real-world natural scenes and the benefits of using our TG-GAN are clear. StyTr², for example, occasionally fails to adequately transfer style on some complex textures e.g. parts with foliage or water reflections, and S2WAT has issue of applying fine detail through multiple window attentions e.g during coastal scenes.

5. CONCLUSION

In this paper, proposed a new framework called TG-GAN, a novel Transformer-Guided Generative Adversarial Network that combines the capability of transformers to model global features with the image-synthesis creativity of GANs for efficient style transfer on a range of artistic and natural scene datasets. Our approach captures long dependency structures efficiently using multi-head self-attention in both the generator and discriminator to enhance the content preservation of the natural scenes and faithfully embed the artistic style. Extensive benchmarking using standard datasets MS-COCO, WikiArt, and Flickr Landscapes show our model achieved the best performance in terms of SSIM, FID, LPIPS, and Style Accuracy over strong baselines including AdaIN, SANet, StyTr², and DualStyleGAN. The proposed TG-GAN model demonstrated competitive inference time while preserving perceptual quality and runtime efficiency, which is important in real-time systems. In the future work, we desire to integrate dynamic style control methods that allow user-guided intensity changes and targeting of semantic regions. Model can also be extended to allow for temporal consistency in video style transfer in conjunction with transformer-based temporal encodes. Additionally, model can be extended to allow for multimodal conditioning using text or audio prompts for interactive and context-aware stylization.

REFERENCES

[1] Mei, L.R., Yao, J.Y., Ge, Y.Y., Wang, Y.W., et al. (2025). A survey of context engineering for large language models. arXiv preprint arXiv:2507.13334. https://doi.org/10.48550/arXiv.2507.13334

[2] Joshi, A., Shet, A.V., Thambi, A.S., R, S. (2023). Quality improvement of image datasets using hashing techniques. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, pp. 18-23. https://doi.org/10.1109/IITCEE57236.2023.10091044

[3] Chen, H.W., Shao, F., Chai, X.L., Jiang, Q.P., Meng, X.C., Ho, Y.S. (2024). Collaborative Learning and style-adaptive pooling network for perceptual evaluation of arbitrary style transfer. IEEE Transactions on Neural Networks and Learning Systems, 35(11): 15387-15401. https://doi.org/10.1109/TNNLS.2023.3286542

[4] Lin, Z., Wang, Z.Z., Chen, H.B., Ma, X.L., Xie, C., Xing, W. (2021). Image style transfer algorithm based on semantic segmentation. IEEE Access, 9: 54518-54529. https://doi.org/10.1109/ACCESS.2021.3054969

[5] Joshi, A., Shet, A.V., Thambi, A.S. (2023). Quality improvement of image datasets using Hashing Techniques. In 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, pp. 18-23. https://doi.org/10.1109/IITCEE57236.2023.10091044

[6] Chen, Y.S., Yuan, Q., Li, Z.Q., Liu, Y.G., Wang, W., Xie, C.P., Wen, X.M., Yu, Q. (2024). UPST-NeRF: Universal photorealistic style transfer of neural radiance fields for 3D scene. IEEE Transactions on Visualization and Computer Graphics, 31(4): 2045-2057. https://doi.org/10.1109/TVCG.2024.3378692

[7] An, J., Huang, S., Song, Y., Dou, D., Liu, W., Luo, J. (2021). ArtFlow: Unbiased image style transfer via reversible neural flows. In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 862-871. https://doi.org/10.1109/CVPR46437.2021.00092

[8] Chen, G., Zhang, G.P., Yang, Z.G., Liu, W.Y. (2023). Multi-scale patch-GAN with edge detection for image inpainting. Applied Intelligence, 53(4): 3917-3932. https://doi.org/10.1007/s10489-022-03577-2

[9] Wang, W.J., Yang, S., Xu, J.Z., Liu, J.Y. (2020). Consistent video style transfer via relaxation and regularization. IEEE Transactions on Image Processing, 29: 9125-9139. https://doi.org/10.1109/TIP.2020.3024018

[10] Liao, Y.S., Huang, C.R. (2022). Semantic context-aware image style transfer. IEEE Transactions on Image Processing, 31: 1911-1923. https://doi.org/10.1109/TIP.2022.3149237

[11] Liu, S.G., Zhu, T. (2022). Structure-guided arbitrary style transfer for artistic image and video. IEEE Transactions on Multimedia, 24: 1299-1312. https://doi.org/10.1109/TMM.2021.3063605

[12] Xu, Z.J., Hou, L.Y., Zhang, J.Q. (2022). IFFMStyle: High-quality image style transfer using invalid feature filter modules. Sensors, 22(16): 6134. https://doi.org/10.3390/s22166134

[13] Ma, Z.Q., Lin, T.W., Li, X., Li, F., He, D.L., Ding, E. (2023). Dual-affinity style embedding network for semantic-aligned image style transfer. IEEE Transactions on Neural Networks and Learning Systems, 34(10): 7404-7417. https://doi.org/10.1109/TNNLS.2022.3143356

[14] Pan, X.D., Zhang, M., Ding, D.Z., Yang, M. (2022). A geometrical perspective on image style transfer with adversarial learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(1): 63-75. https://doi.org/10.1109/TPAMI.2020.3011143

[15] Qu, Y., Shao, Z.Z., Qi, H.R. (2022). Non-local representation based mutual affine-transfer network for photorealistic stylization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 7046-7061. https://doi.org/10.1109/TPAMI.2021.3095948

[16] Singh, A., Jaiswal, V., Joshi, G., Sanjeeve, A., Gite, S., Kotecha, K. (2021). Neural style transfer: A critical review. IEEE Access, 9: 131583-131613. https://doi.org/10.1109/ACCESS.2021.3112996

[17] Wang, X.H., Wang, W.J., Yang, S., Liu, J.Y. (2022). CLAST: Contrastive learning for arbitrary style transfer. IEEE Transactions on Image Processing, 31: 6761-6772. https://doi.org/10.1109/TIP.2022.3215899

[18] Chen, H.B., Wang, Z.J., Zhao, L., Li, J., Yang, J. (2025). TRTST: Arbitrary high-quality text-guided style transfer with transformers. IEEE Transactions on Image Processing, 34: 759-771. https://doi.org/10.1109/TIP.2025.3530822

[19] Hua, S.H., Zhang, D.D. (2021). Multi-attention network for arbitrary style transfer. In Neural Information Processing, pp. 390-401. https://doi.org/10.1007/978-3-030-92273-3_32

[20] Chen, Y.W., Chen, R., Lei, J.B., Zhang, Y.B., Jia, K. (2022). TANGO: Text-driven photorealistic and robust 3D stylization via lighting decomposition. In Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22), NY, USA, pp. 30923-30936. https://dl.acm.org/doi/10.5555/3600270.3602512.

[21] Deng, J.H., Zhang, X.Y., Li, W., Duan, L.X. (2022). Cross-domain Detection Transformer based on spatial-aware and semantic-aware token alignment. arXiv preprint arXiv:2206.00222. https://doi.org/10.48550/arXiv.2206.00222

[22] Park, J., Kim, Y. (2021). Styleformer: Transformer based generative adversarial networks with style vector. arXiv preprint arXiv:2106.07023. https://doi.org/10.48550/arXiv.2106.07023

[23] Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T. (2023). StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515. https://doi.org/10.48550/arXiv.2301.09515

[24] Cui, S.Y., Hui, B. (2024). Dual-dependency attention transformer for fine-grained visual classification. Sensors, 24(7): 2337. https://doi.org/10.3390/s24072337

[25] Chen, Y.M., Xu, J.H., An, Z.L., Zhuang, F.Z. (2024). Multi-scale conditional reconstruction generative adversarial network. Image and Vision Computing, 141: 104885. https://doi.org/10.1016/j.imavis.2023.104885

[26] Zhang, B.W., Gu, S.Y., Zhang, B., Bao, J.M., Chen, D., Wen, F., Wang, Y., Guo, B.N. (2021). StyleSwin: Transformer-based GAN for high-resolution image generation. arXiv preprint arXiv:2112.10762. https://doi.org/10.48550/arXiv.2112.10762

[27] Bi, Y., Abrol, A., Jia, S., Sui, J., Calhoun, V.D. (2024). Gray matters: ViT-GAN framework for identifying schizophrenia biomarkers linking structural MRI and functional network connectivity. NeuroImage, 297: 120674. https://doi.org/10.1016/j.neuroimage.2024.120674

[28] Huang, Y.H., He, Y., Yuan, Y.J., Lai, Y.K., Gao, L. (2022). StylizedNeRF: Consistent 3D scene stylization as stylized NeRF via 2D-3D mutual learning. arXiv preprint arXiv:2205.12183. https://doi.org/10.48550/arXiv.2205.12183

[29] Cho, H., Lee, J., Chang, S., Jeong, Y. (2024). One-shot structure-aware stylized image synthesis. arXiv preprint arXiv:2402.17275. https://doi.org/10.48550/arXiv.2402.17275

[30] Huang, X., Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. arXiv preprint arXiv:1703.06868. https://doi.org/10.48550/arXiv.1703.06868

[31] Park, D.Y., Lee, K.H. (2018). Arbitrary style transfer with style-attentional networks. arXiv preprint arXiv:1812.02342. https://doi.org/10.48550/arXiv.1812.02342

[32] Kolkin, N., Salavon, J., Shakhnarovich, G. (2019). Style transfer by relaxed optimal transport and self-similarity. arXiv preprint arXiv:1904.12785. https://doi.org/10.48550/arXiv.1904.12785

[33] Deng, Y.Y., Tang, F., Dong, W.M., Ma, C.Y., Pan, X.J., Wang, L., Xu, C.S. (2021). StyTr^2: Image Style transfer with transformers. arXiv preprint arXiv:2105.14576. https://doi.org/10.48550/arXiv.2105.14576

[34] Zhang, C.Y., Xu, X.G., Wang, L., Dai, Z.Y., Yang, J. (2022). S2WAT: Image style transfer via hierarchical vision transformer using strips window attention. arXiv preprint arXiv:2210.12381. https://doi.org/10.48550/arXiv.2210.12381

[35] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014). Microsoft COCO: Common objects in context. In Computer Vision – ECCV 2014, pp. 740-55. https://doi.org/10.1007/978-3-319-10602-1_48

[36] Saleh, B., Elgammal, A. (2015). Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855. https://doi.org/10.48550/arXiv.1505.00855

[37] Young, P., Lai, A., Hodosh, M., Hockenmaier, J. (2014). 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. https://doi.org/10.1162/tacl_a_00166