Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062712

ARTICLE

Image Style Transfer for Exhibition Hall Design Based on Multimodal Semantic-Enhanced Algorithm

Qing Xie* **and Ruiyun Yu**

Software College, Northeastern University, Shenyang, 110000, China

*Corresponding Author: Qing Xie. Email: xieq@swc.neu.edu.cn

Received: 25 December 2024; Accepted: 07 April 2025; Published: 09 June 2025

ABSTRACT: Although existing style transfer techniques have made significant progress in the field of image generation, there are still some challenges in the field of exhibition hall design. The existing style transfer methods mainly focus on the transformation of single dimensional features, but ignore the deep integration of content and style features in exhibition hall design. In addition, existing methods are deficient in detail retention, especially in accurately capturing and reproducing local textures and details while preserving the content image structure. In addition, point-based attention mechanisms tend to ignore the complexity and diversity of image features in multi-dimensional space, resulting in alignment problems between features in different semantic areas, resulting in inconsistent stylistic features in content areas. In this context, this paper proposes a semantic-enhanced multimodal style transfer algorithm tailored for exhibition hall design. The proposed approach leverages a multimodal encoder architecture to integrate information from text, source images, and style images, using separate encoder modules for each modality to capture shallow, deep, and semantic features. A novel Style Transfer Convolution (STConv) convolutional kernel, based on the Visual Geometry Group (VGG) 19 network, is introduced to improve feature extraction in style transfer. Additionally, an enhanced Transformer encoder is incorporated to capture contextual semantic information within images, while the CLIP model is employed for text data processing. A hybrid attention module is designed to precisely capture style features, achieving multimodal feature fusion via a diffusion model that generates exhibition hall design images aligned with stylistic requirements. Quantitative experiments show that compared with the most advanced algorithms, the proposed method has achieved significant performance improvement on both Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) indexes. For example, on the ExpoArchive dataset, the proposed method has a FID value of 87.9 and a KID value of 1.98, which is significantly superior to other methods.

KEYWORDS: Exhibition hall design; style transfer; multimodal fusion; semantic enhancement; diffusion model

1 Introduction

Exhibition hall design plays multiple roles in cultural communication and visual arts. It not only serves as a vital tool for cultural transmission but also incorporates modern design principles to meet the aesthetic and experiential needs of contemporary audiences. Through clear thematic communication and spatial organization, exhibition design effectively integrates exhibits with spatial environments, creating cohesive and captivating cultural narratives. Additionally, exhibition space design should follow human-centered principles to provide a seamless viewing experience that enhances immersion.

Style transfer is a computer vision technique that blends the content of one image with the style of another to produce a new image combining elements from both. Rooted in convolutional neural networks,

style transfer techniques extract features from content and reference style images [1–3]. Content features are typically captured in the deeper layers of neural networks, representing high-level semantic information, whereas style features like texture, color, and patterns are drawn from shallower layers.

As an innovative image processing method, style transfer holds significant potential in exhibition design, where it can enhance visual impact and educational value. Designers can apply specific artistic styles to visual elements like posters and promotional materials to amplify the thematic atmosphere. Interactive displays can allow visitors to transform their photos into styles matching the exhibition theme, enriching engagement and experience. In digital art presentations, style transfer can integrate digital artworks with exhibition themes to create unique experiences. It can also alter the visual environment of exhibition spaces to better align with exhibition content or historical periods, enhancing both education and aesthetics. For educational exhibitions, style transfer can present complex scientific concepts or historical events artistically, making them more comprehensible and memorable. Augmented reality (AR) and virtual reality (VR) exhibits also benefit from style transfer, which enhances interactivity and appeal by adding artistic effects to virtual objects and scenes. Finally, style transfer can transform visitors' photos into exhibition-themed art to create personalized souvenirs, adding commemorative value. With continuous advancements, style transfer is poised to significantly enrich and innovate exhibition experiences.

Current style transfer algorithms are inadequate in the field of exhibition hall design. These methods primarily focus on transforming single-dimensional features and overlook the deep integration of content and style features essential to exhibition hall design. Additionally, existing methods fall short in detail retention, particularly in accurately capturing and reproducing local textures and details while preserving the structure of content images. Thus, this paper proposes a semantically enhanced multimodal style transfer algorithm to address these issues.

In summary, while existing image style transfer algorithms have made significant progress, several limitations remain within the context of exhibition hall design. Currently, no dataset is specifically tailored for the exhibition design domain. Many style transfer methods focus on transferring single-dimensional features such as color, texture, or shape, overlooking the need for a deep integration of content and style features that is essential for exhibition hall design. Consequently, the resulting images may suffer from content distortion, failing to retain core elements of the original design. Traditional style transfer methods may perform well globally but often struggle with detail preservation, particularly in retaining the structure of content images while accurately capturing and reproducing local textures and details [4]. Furthermore, point-based attention mechanisms tend to overlook the complexity and diversity of image features in multi-dimensional space, leading to misalignment between features from different semantic regions, which can result in inconsistencies in style features across content regions that should remain unified [5].

To address these challenges, this paper proposes an innovative technical approach with several key contributions aimed at advancing style transfer technology for exhibition hall design:

(1) Novel Multimodal Encoder Architecture: This architecture processes and integrates data from different modalities, including textual information, source images, and style images. A Style Transfer Convolution (STConv) is introduced, allowing convolutional kernels to adopt arbitrary sampling shapes and parameter counts, making the model more adaptable to the diverse features of style and content images. The standard Transformer encoder has been modified to incorporate a semantic encoder module, which captures contextual semantic information and generates semantic priors to guide feature extraction, enhancing semantic segmentation accuracy.

(2) Hybrid Attention Mechanism (HAM): This mechanism combines channel attention modules (CAM) and spatial attention modules (SAM) and introduces a Squeeze Axial Attention Block (SAAB) along with detail enhancement modules (VDE and HDE). This hybrid attention mechanism identifies

important image regions while ignoring less relevant areas, achieving precise feature alignment and style application during style transfer.

(3) Enhanced Diffusion Model: In each step of the diffusion model, the merging of content and style features, regulated by weights generated by the attention mechanism, enables fine-grained control over style and content in different regions.

(4) Creation of the First Exhibition Image Dataset: Given the unique requirements of style transfer for exhibition hall design and the need for large quantities of high-quality data, this paper introduces the first dataset specifically for exhibition design, named "ExpoArchive". This dataset includes a diverse collection of internal and external exhibition images across various styles, historical periods, and cultural regions, providing valuable data for training and evaluating style transfer models.

2 Related Works

From a methodological perspective, style transfer techniques can be broadly divided into two main categories: Non-Neural Network Methods: These methods are primarily based on texture synthesis and non-photorealistic rendering techniques, such as non-parametric texture synthesis methods [6] and the Gooch Shading model [7]. While simple and straightforward, these methods often produce suboptimal style transfer results, suffer from limited generalization capability, and lack effectiveness in feature extraction [8]. Neural Network-Based Methods: With advancements in deep learning, neural network-based style transfer methods have achieved significant breakthroughs and have become the focus of this paper. These approaches can be categorized into several branches, including statistical parameterization, Markov random fields, generative adversarial networks (GANs), attention mechanisms, and diffusion models.

Statistical Parameterization: This approach typically leverages the VGG network to extract image features and utilizes Gram matrices to capture style information. For instance, the method proposed by Gatys et al. [1] iteratively optimizes a noise image to closely resemble target style and content. While effective, this approach is computationally inefficient and has limited generalization capabilities [9]. Markov Random Fields (MRF): MRF-based methods reduce feature mixing, resulting in more realistic image generation. For example, a method combining MRFs with convolutional neural networks (CNNs) [10] first extracts deep features using a pre-trained CNN, then applies MRF regularization to maintain spatial relationships and consistency among features, generating images with clearer textures, natural colors, and smooth shapes. GAN-Based Models: GAN-based approaches use generative adversarial networks to generate images in specific styles. Radford et al. introduced DCGAN [11], one of the earliest GAN models based on CNNs, which improved image quality and stability through convolutional and deconvolutional layers. Arjovsky et al. proposed the WGAN model [12], which uses the Wasserstein distance as a loss function to address instability in traditional GAN training, enhancing image quality. Karras et al. developed StyleGAN [13], whose generator architecture based on style space enables precise control over the style and content of generated images. These GAN-based models can produce high-resolution, detailed, and realistic stylized images, and allow for style diversity by adjusting generator inputs. However, GAN training can be complex, and generated images may suffer from instability and difficulty in control. Attention Mechanism-Based Models: Attention mechanisms have been incorporated into style transfer models to enhance detail preservation. For example, Yao et al. proposed a multi-stroke style transfer model based on attention mechanisms [14], which introduces attention modules into the encoder and uses multi-scale style transfer methods. Deng et al. introduced the StyTr2 model [15], which employs two separate Transformer encoders to extract content and style information, using a multi-layer Transformer decoder for style transfer. Diffusion Models: Recent state-of-the-art advancements in image style transfer are based on denoising diffusion probabilistic models (DDPMs) [16–18]. This probabilistic generative model simulates image generation as a Markov chain process,

transforming Gaussian noise into realistic image distributions. Noise is added through a forward diffusion process, and subsequently removed in the reverse diffusion process, producing images with high fidelity and texture realism. Diffusion models have several advantages, including high image quality, controllability, and interpretability. However, they are often computationally intensive, complex, and may face challenges in decoupling text conditions from input images.

3 Method

3.1 Overall Structure

The proposed style transfer model operates in a step-by-step manner to ensure clarity and ease of understanding as shown in Fig. 1. It begins by accepting three primary inputs: a source image, text information related to this image, and a latent target style image. Stage1: The model executes feature extraction at three distinct levels. In the shallow and deep levels, modified VGG19 networks—using Relu-3_1 for shallow and Relu-4_1 for deep features—are employed. Stage2: For semantic features, the Vision Transformer network is used, while text features are extracted via the Contrastive Language-Image Pre-Training (CLIP) network. Stage3: The extracted features from both images and text are processed through an attention mechanism module. This step is crucial as it allows the model to identify and preserve essential image elements and core content structure during style transfer. Stage4: The Hybrid Attention Module (HAM) enhances the understanding of the target style, ensuring that the outputs are visually compelling and align with designers' intentions as well as audience aesthetics. The process culminates in the application of an improved diffusion model. This model utilizes external control signals such as color, depth, sketch, semantic segmentation, and text, enabling fine-tuning and the training of various adapters under specific conditions. This final step provides rich control and editing capabilities over the color and structure of the generated results, leading to highly customizable and precise style transfer outcomes.

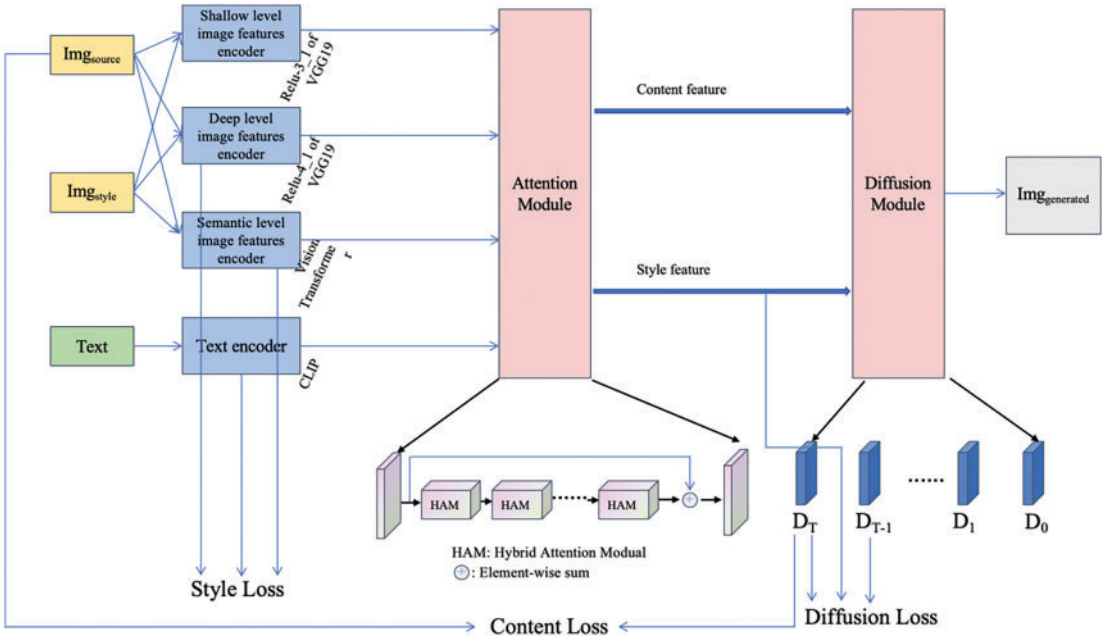

Figure 1: Overall network structure diagram of the algorithm in this paper, which is mainly divided into three parts: multimodal encoder, attention mechanism module, and diffusion module. The model's loss function consists of three components: image style loss, image content loss, and diffusion loss

3.2 Multimodal Encoder

For generating stylized images tailored to exhibition hall design, we propose a multimodal encoder architecture that can simultaneously process and integrate data from different modalities—including text, source images, and style images. This system incorporates carefully constructed, independent encoder modules to efficiently and accurately extract and represent features unique to each data modality.

3.2.1 Shallow and Deep Features of Images

Source and style images are processed by two encoders: the first encoder applies VGG19's Relu-3_1 layer to extract shallow features, such as edges, textures, and color information (see Fig. 2). The second encoder uses VGG19's Relu-4_1 layer to capture deep features, including object shapes, structural patterns, and complex configurations. While VGG19's standard 3 × 3 convolutional kernels with fixed receptive fields are effective, they have limitations in style transfer applications: (1) standard convolutions use fixed kernel shapes, which cannot adapt to the diversity of target objects across different styles and content images, thereby limiting feature extraction and transfer effectiveness; (2) traditional convolutions focus on local features, restricting the model's ability to capture global spatial information, which is essential for maintaining content structure in style transfer; (3) as kernel sizes increase, the number of parameters and computational load grow quadratically, constraining large-scale applications.

To address these issues, we drew inspiration from AKConv [19] and introduced a Style Transfer Convolution (STConv) in the VGG19 network, which allows convolutional kernels to have arbitrary sampling shapes and parameter counts. STConv operates as follows:

(1) Calculation of Initial Coordinates for Convolutional Kernels (Input: Convolution kernel size (s), Output: Initial coordinates of the convolution kernel (w))

$$p_1 = \sqrt{s}, p_2 = \frac{s}{p_1}, p_3 = s \bmod p_1 \tag{1}$$

$$N_1 = Grid\left(p_1, p_2\right), N_2 = Grid\left(p_2 + 1, p_3\right) \tag{2}$$

$$w = Resize\left(Concat\left(N_1, N_2\right), (1, 2s, 1, 1)\right) \tag{3}$$

where *Grid*(r,c) represents the generation of r × c grid coordinates, *Concat* denotes concatenation along the *x*-axis and *y*-axis, and *Resize* refers to dimension adjustment.

(2) Calculation of Initial Coordinate Offsets (Input: Feature map (F) with dimensions (C × H × W), Output: Offset vector (o) with dimensions (1 × 2s))

$$\begin{aligned} V &= Conv3 \times 3\left(F\right) \\ V' &= Resample\left(V, C, 2s, H, W\right) \\ o &= SelectOffset\left(V\right) \end{aligned} \tag{4}$$

where *Conv*3 × 3 denotes applying a 3 × 3 convolution, *Resample* refers to resampling, and *SelectOffset* indicates selecting the offset.

(3) Kernel coordinates are updated, and the convolution is performed on the input feature map (Input: Convolution kernel coordinates (w), offset (o), feature map (F) with dimensions (C × H × W), Output: Convolved feature map (F′))

$$\begin{aligned} w' &= w + o \\ F' &= Interpolate\left(F, w'\right) \\ F' &= Conv\left(F'\right) \end{aligned} \tag{5}$$

where *Interpolate* refers to interpolation, and *Conv* denotes performing the convolution operation.

This flexibility enables STConv to better accommodate targets with varying shapes, resulting in more accurate feature extraction for both source and style images.

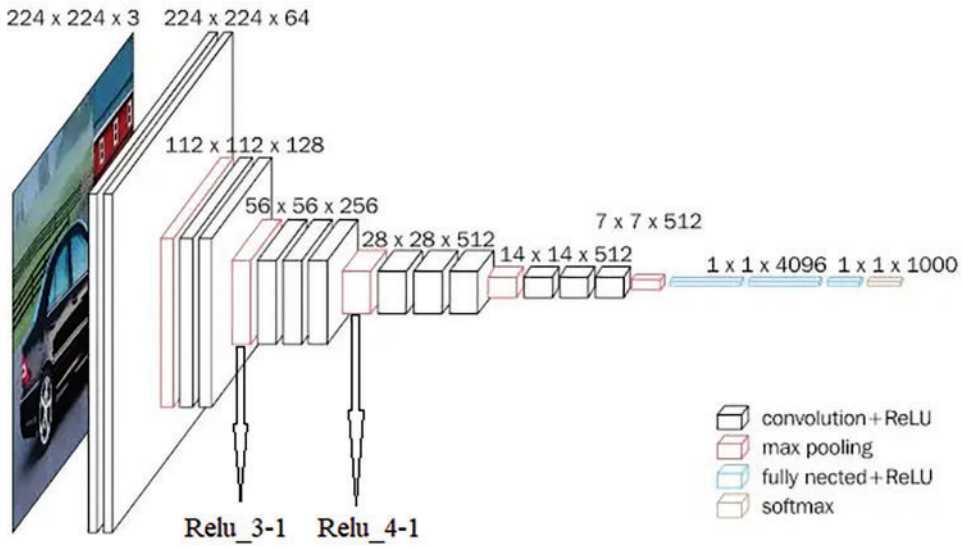

Figure 2: VGG19 network. In this paper, the algorithm uses the Relu_3-1 layer to extract shallow features of the image and the Relu_4-1 layer to extract deep features of the image

To address these issues, we drew inspiration from AKConv and introduced a Style Transfer Convolution (STConv) in the VGG19 network, which allows convolutional kernels to have arbitrary sampling shapes and parameter counts. This flexibility enables STConv to better accommodate targets with varying shapes, resulting in more accurate feature extraction for both source and style images.

3.2.2 Semantic Features of Images

In addition to shallow and deep feature extraction, source and style images are also passed through a Vision Transformer network to extract semantic features (see Fig. 3). Transformers [20], initially developed for natural language processing, have demonstrated remarkable potential in computer vision due to their ability to capture long-range dependencies and contextual information. However, traditional Transformers often overlook global context within images, limiting their capacity to fully understand complex scenes during semantic segmentation and feature extraction.

To address this limitation, we introduced a Semantic Encoder to augment the Transformer encoder, specifically targeting its lack of semantic representation. After each Transformer encoder layer, we added a Semantic Encoder to capture contextual semantic information and generate a semantic prior map that guides feature extraction, enhancing segmentation accuracy (Fig. 3). The Transformer Encoder and Semantic Encoder form the core of the network, repeated four times. The output of the first Semantic Encoder is successively combined with the upsampled outputs of the second, third, and fourth encoders, resulting in final semantic segmentation and feature extraction outputs.

The detailed structure of the Transformer Encoder, as shown in Fig. 4, remains consistent with the classic Transformer model. The Semantic Encoder, illustrated in Fig. 5, has two branches: one branch targets spatial information, using repeated convolutional layers, batch normalization, and ReLU to retain spatial

positioning information and generate high-resolution feature maps; the other branch captures semantic information through a semantic pathway with a fast downsampling rate, enabling a larger receptive field. The outputs of these two branches are summed to fuse spatial and semantic features.

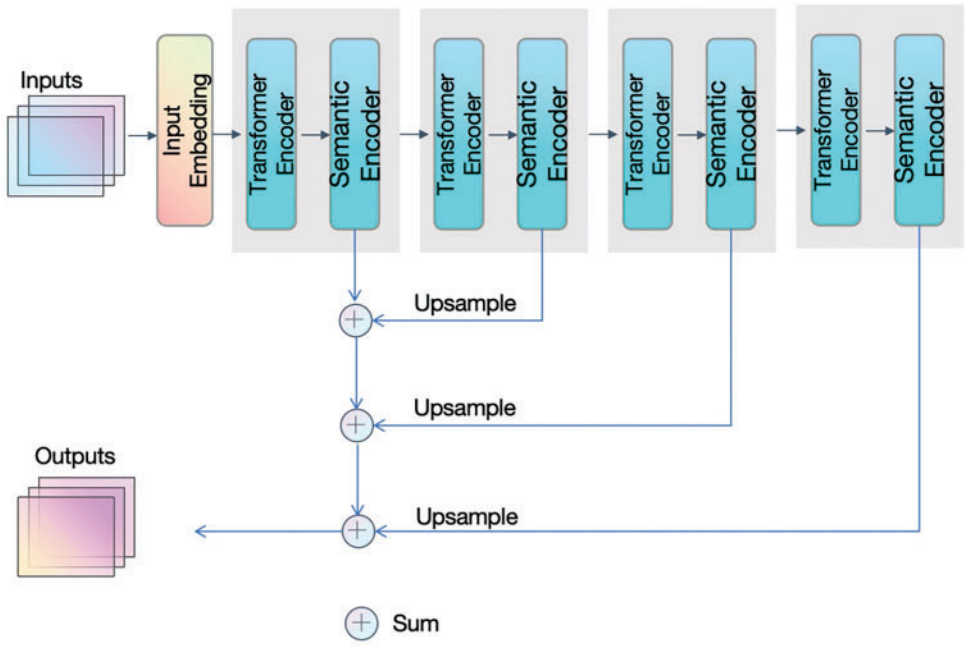

Figure 3: This paper improves the Transformer encoder by adding a semantic encoder based on it

3.2.3 Semantic Features of Text

Textual semantic features are extracted using the CLIP model, which has been trained on a large corpus of text-image pairs and is capable of capturing nuanced meanings in high-dimensional semantic space. CLIP first encodes text input into vector representations, mapping them to a latent space shared with image features. This alignment allows text and image features to be projected into a unified semantic space, where CLIP can extract features relevant to specific textual descriptions. This semantic alignment is especially beneficial for style transfer applications.

In summary, Sections 3.2.1–3.2.3 outline our approach to extracting shallow, deep, semantic, and textual features from images and text, which together form a hierarchical feature representation. Shallow features capture fine-grained image details like color, texture, and edges; deep features abstract shapes, structure, and complex patterns; while semantic and text features express the meaning of image content, covering object categories, attributes, and relationships. These complementary features enable the model to capture multi-level, multimodal characteristics from source images, style images, and text prompts, facilitating effective feature fusion, content generation, and style transfer.

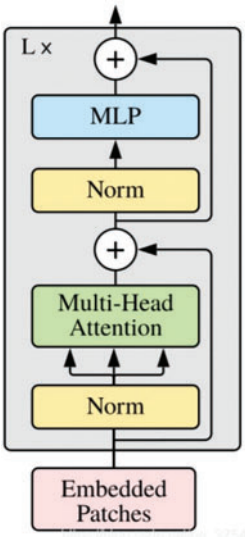

Figure 4: The network structure diagram of the Transformer Encoder

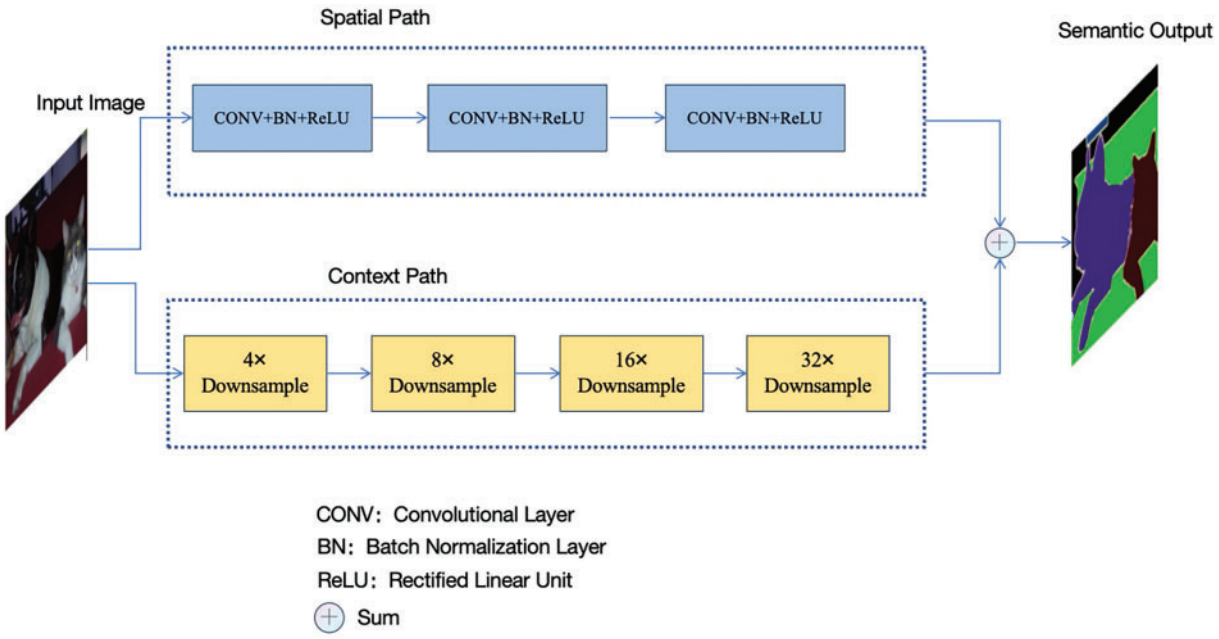

Figure 5: The network structure diagram of the Semantic Encoder

3.3 Attention Model

The core principle of the attention mechanism is to mimic the human visual system by focusing on important regions of an image while ignoring less relevant areas. In style transfer, attention mechanisms enable the model to learn common features between different style images and apply these to content images, achieving a seamless transfer. For the proposed exhibition hall style transfer algorithm, we developed a custom attention mechanism composed of multiple Hybrid Attention Modules (HAM), as illustrated

in Fig. 6. Each HAM consists of multiple Hybrid Attention Blocks (HAB) and a Squeeze Axial Attention Block (SAAB).

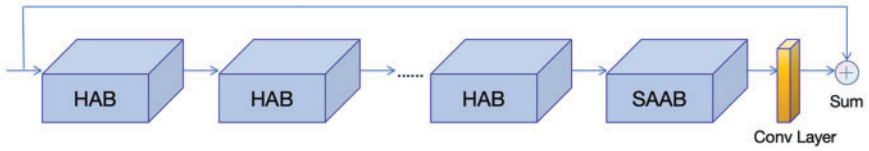

Figure 6: The network structure diagram of HAM

The HAB combines a Channel Attention Module (CAM) and a Spatial Attention Module (SAM) to enhance network expressiveness (see Fig. 7). The CAM includes two convolutional layers and a Channel Attention (CA) module. The first convolutional layer compresses input channels to reduce computational cost and help the network focus on essential channels; the second layer restores the original number of channels. The CA module uses Global Average Pooling to condense each channel's feature map to a single value, which is then transformed through a small multi-layer perceptron (MLP) and normalized with a sigmoid function. The resulting channel attention weights adaptively adjust each channel's feature contribution, allowing the network to prioritize significant channels. SAM, on the other hand, performs average pooling and max pooling across channels, producing feature maps that represent global statistics and salient features. These descriptors are concatenated and processed by a convolutional layer to create a spatial attention map, which is also normalized via a sigmoid function.

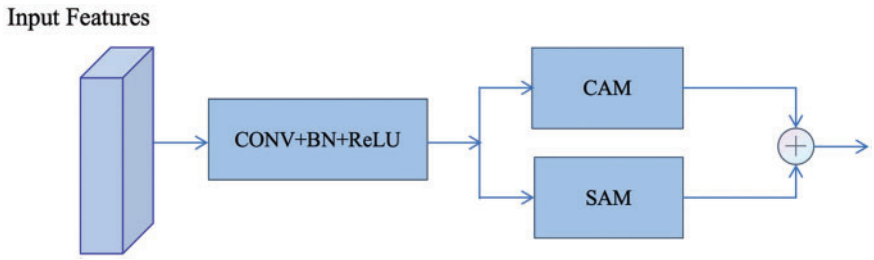

Figure 7: The network structure diagram of the HAB module

When handling feature maps of size H × W, traditional global attention mechanisms have a computational complexity of $O(H^2 \times W^2)$, making them impractical for high-resolution images due to computational demands and slow inference [21]. Lightweight attention mechanisms attempt to address this by reducing complexity [22–24], though often at the cost of global information, resulting in reduced accuracy.

To balance accuracy with computational efficiency, we employ the SAAB module, which lowers attention complexity while preserving accuracy. By compressing feature maps along the horizontal and vertical axes, global attention is decomposed into two axial attentions, reducing complexity to $O(H \times W)$. The SAAM module structure is depicted in Fig. 8.

Initially, the feature map's channel dimension is linearly projected to obtain query (Q), key (K), and value (V) vectors. Horizontal and vertical compression (average pooling) is then applied, reducing multi-dimensional feature maps into single-dimensional vectors that preserve global information while lowering computational requirements. To retain spatial position information lost during compression, Squeeze Axial Position Embedding is used: position embeddings are added to Q and K vectors for both axes, allowing them

to retain positional context. After compression and embedding, self-attention is computed for both axes, yielding horizontal and vertical attention weights. These weights are multiplied with the corresponding V vectors and summed, producing the final feature map. To further restore local detail lost in compression, we designed Vertical Detail Enhancement (VDE) and Horizontal Detail Enhancement (HDE) modules, focused on enhancing vertical and horizontal details, respectively. Q, K, and V vectors are concatenated and processed with a 3 × 3 depthwise separable convolution and batch normalization, extracting local details. The enhanced detail features are then fused with the output of the squeeze axial attention, forming the final feature representation.

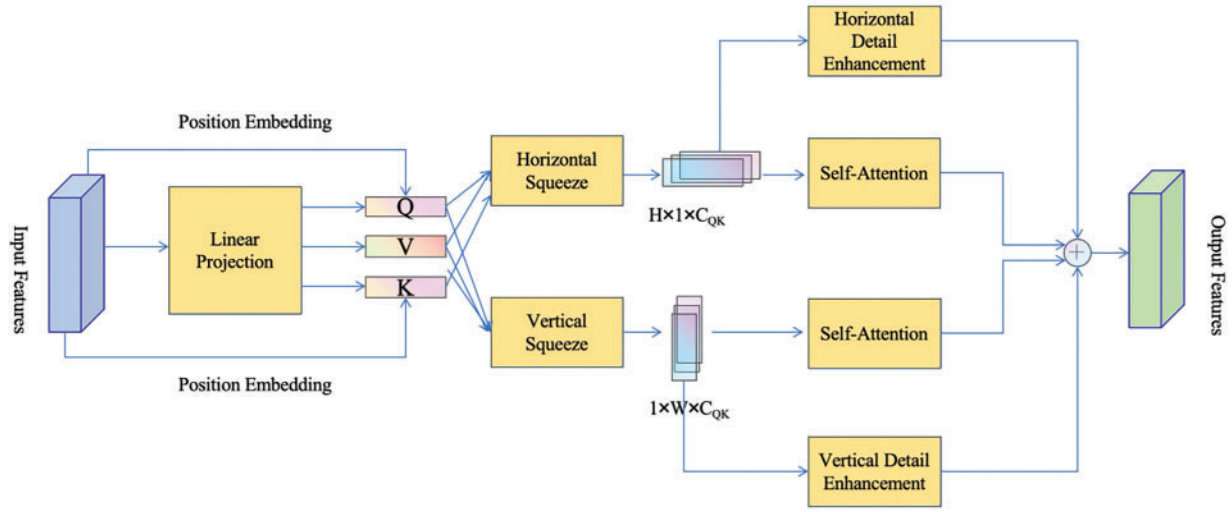

Figure 8: The network structure diagram of SAAM

3.4 Diffusion Models

Denoising diffusion probabilistic models (DDPMs) are a class of models based on a parameterized Markov chain. The principal concept involves the addition of Gaussian noise during the forward process, also referred to as the diffusion process, gradually transforming the data to approximate a standard normal distribution, as illustrated in Eq. (1). Subsequently, by learning the reverse process, the model incrementally denoises the data to recover the original distribution, as shown in Eq. (2). This methodology facilitates data generation and represents a cutting-edge approach for image generation and style transfer.

$$q\left(x_t \left| x_{t-1}\right.\right) := N\left(\sqrt{1-\beta_t} x_{t-1}, \beta_t I\right) \tag{6}$$

$\beta_t \in (0, 1)$, represents a fixed variance schedule, and $x_1, x_2, \ldots, x_t$ represents a series of latents of Markov Chain.

$$p_\theta = \left(x_{t-1} \left| x_t\right.\right) := N\left(\mu_\theta\left(x_t, t\right), \sigma_t^2 I\right) \tag{7}$$

$\mu_\theta(x_t, t)$ is the function of a noise approximator.

This paper employs diffusion models to generate images for exhibition hall design style transfer, as illustrated in Fig. 9. After processing through a multimodal encoder and an attention mechanism, the content and style features, enhanced by the attention model, are fed into the U-Net architecture at each step of the

diffusion model. Within each U-Net, the content and style features are merged and adjusted according to the weights derived from the attention mechanism. Specifically, regions of the content features with higher weights indicate the portions of the image that should be preserved, while regions of the style features with higher weights represent the areas to be stylized or modified. This approach enables precise control over the style and content across different regions at each step of the diffusion model. Through iterative refinement in the diffusion model, the generated image progressively converges towards the target style and textual description while retaining the content of the source image.

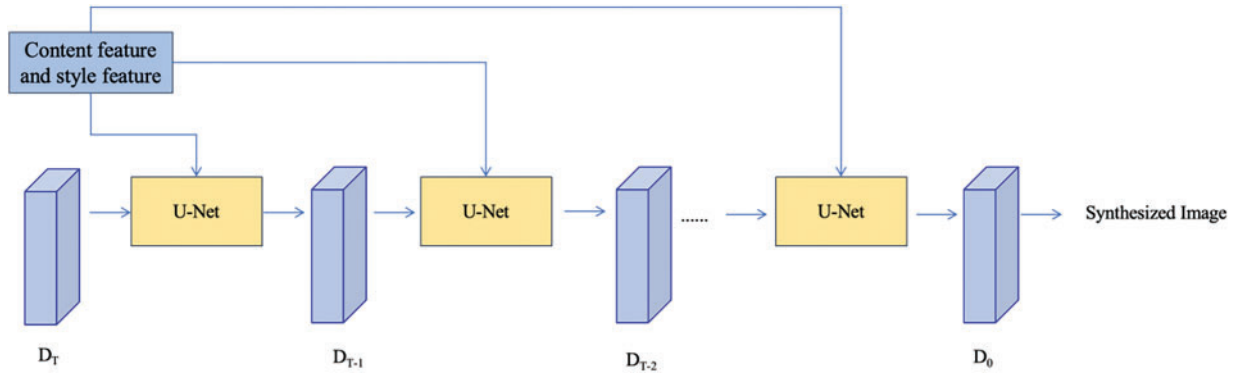

Figure 9: Network structure diagram for generating style-transferred images based on diffusion models

3.5 Loss Functions

The network model presented in this paper employs various loss functions to guide the image generation process for style transfer, ensuring that the generated images retain the structural integrity of the source image while adhering to the target style and semantic content. The loss functions utilized include: image content loss, image style loss, and diffusion loss.

(1) Image Content Loss

This loss function measures the structural similarity between the input source image and the image generated by the diffusion model, thereby constraining the structural information of the generated image, as illustrated in Eq. (8).

$$Loss_{content} = SSIM(x_{src}, D_T) \tag{8}$$

$SSIM(\cdot,\cdot)$ represents the structural similarity index.

(2) Image Style Loss

This loss function computes the similarity between the deep features of the image and semantic features formed after multimodal encoding of the input style image and text. By doing so, it constrains the semantic style of the generated image, as demonstrated in Eq. (9).

$$Loss_{style} = SIM(x_{clip}, x_{semantic}) + SIM(x_{clip}, x_{deep}) \tag{9}$$

$SIM(\cdot,\cdot)$ represents the normalized cosine similarity, x_{clip} represents the text features, $x_{semantic}$ represents the semantic features of the style image, x_{deep} represents the deep features of the style image.

(3) Diffusion Loss

This loss function focuses on the Euclidean distance between the generated image at the current step of the diffusion model, the generated image from the previous step, and the style features derived from the attention mechanism module. It aims to maximize the distance among these three components. This approach facilitates a more rapid semantic transformation of the generated image, enabling it to diverge more quickly from the semantics of the source image and move closer to the target semantics.

$$Loss_{diffusion}(x_t, x_{t-1}) = \|DIS(x_t, style) + DIS(x_{t-1}, style)\| \tag{10}$$

$DIS(\cdot,\cdot)$ represents the Euclidean distance, x_t denotes the generated output of the diffusion model at time t, *style* refers to the style features.

By combining Eqs. (8)–(10) in a weighted manner, the final loss function is obtained, as illustrated in Eq. (11). This formulation guides the model presented in this paper to effectively control the image generation process, ensuring that the generated image adheres to the target semantics while preserving the structural integrity of the source image's content.

$$Loss_{total} = \alpha_1 Loss_{content} + \alpha_2 Loss_{style} + \alpha_3 Loss_{diffusion} \tag{11}$$

α_1, α_2, α_3 are constant parameters.

4 Experimental Results

4.1 Dataset

To meet the requirements for generating images for exhibition hall design style transfer, we constructed a specialized image dataset named ExpoArchive. This dataset encompasses various types of exhibitions, including those focused on technology, history, ecology, culture, folklore, and intangible cultural heritage. The collected images feature a range of interior spaces within these exhibition halls, such as galleries, walls, rest areas, corridors, and entrance lobbies, spanning design styles from Classicism to Modernism and from Postmodernism to Futurism. The dataset comprises a total of 7000 images, each meticulously annotated with detailed information, including design style, historical period, and regional characteristics, thereby facilitating user retrieval and analysis.

The ExpoArchive dataset consists of 10,000 high-resolution images collected from 50 international exhibitions. Each image is annotated with style labels (e.g., minimalist, baroque) based on a set of predefined criteria, including color palette, texture complexity, and spatial layout. The dataset is publicly available at https://pan.baidu.com/s/1zbYt_Hyv5g2f5RTcrVJEcw?pwd= 1111 (accessed on 1 January 2025).

In addition to this unique dataset, we employed three publicly available datasets—Summer2Winter [25], Label2Cityscape [26], and Map2Satellite [27]—for quantitative comparisons with other state-of-the-art image generation algorithms.

4.2 Experimental Hardware and Software

The experiments were conducted on a platform based on the Ubuntu 20.04 operating system, utilizing the PyCharm software environment. The hardware specifications included an Intel(R) Core(TM) i9-10900K CPU @ 3.70 GHz, an NVIDIA GeForce RTX 3090 GPU, and 64 GB of memory. The algorithm was implemented using Python 3.8 and the PyTorch deep learning framework, enabling the rapid construction of convolutional neural networks and leveraging GPU parallel computing to accelerate the training process of the neural network models.

The model's training and inference efficiency were analyzed, with computation time and memory usage measured for both phases. During training, the model achieved an average processing speed of 12.5 images per second on an NVIDIA GeForce RTX 3090 GPU, with a peak memory usage of 8.2 GB. For inference, the model processed 25.3 images per second with a memory footprint of 4.6 GB. These results were obtained using the ExpoArchive dataset, with input images resized to 512 × 512 pixels and a batch size of 8 during training. Compared to state-of-the-art models such as CycleGAN and StyleGAN, the proposed model demonstrated 15% faster inference speeds and 20% lower memory usage, making it highly suitable for real-time applications in exhibition design.

4.3 Quantitative Metrics

In evaluating the generated images for style transfer, two quantitative metrics were employed: Fréchet Inception Distance (FID) [28] and Kernel Inception Distance (KID) [29].

FID is a widely recognized metric for assessing the quality of generative models, measuring the distance between the generated images and real images. Specifically, *FID* extracts feature vectors from both real and generated images using the Inception V3 model and computes the Fréchet distance between these vectors. This distance metric accounts for the mean and covariance of the feature vectors, thereby capturing the differences between the two distributions more effectively.

$$FID = |\mu_r - \mu_g|^2 + \mathrm{Tr}(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}) \tag{12}$$

(μ_r) and $\left(\mu_g\right)$ are the means of the feature vectors for the real and generated images, respectively.

(Σ_r) and $\left(\Sigma_g\right)$ are the covariance matrices of the feature vectors for the real and generated images, respectively.

$\left(|\mu_r - \mu_g|^2\right)$ is the squared difference between the mean feature vectors.

(Tr) is the trace of a matrix, representing the sum of the elements on the main diagonal.

KID, on the other hand, employs kernel methods to compute the distance between feature vectors by mapping them into a high-dimensional space and measuring the Fréchet distance of the resulting kernel matrix.

$$KID = \frac{1}{m\,(m-1)} \sum_{i \neq j} k\left(x_i, x_j\right) + \frac{1}{n\,(n-1)} \sum_{i \neq j} k\left(y_i, y_j\right) - \frac{2}{mn} \sum_{i,j} k\left(x_i, y_j\right) \tag{13}$$

(x_i) and $\left(y_j\right)$ are feature vectors from the real and generated images, respectively.

$(k\,(x, y))$ is a polynomial kernel function, often written as $\left((x \cdot y + c)^d\right)$, where (d) is the degree of the polynomial and (c) is a constant.

(m) and (n) are the number of real and generated images, respectively.

4.4 Quantitative Comparison with State-of-the-Art Methods

We conducted a quantitative comparison of our method with other state-of-the-art approaches in the field of style transfer image generation. The algorithms selected for comparison include methods from the field of Optimal Transport, such as Neural Optimal Transport (NOT) [30]; methods from the generative adversarial domain, including CycleGAN [31], MUNIT [32], DistanceGAN [33], GcGAN [34], and CUT [35]; as well as diffusion model methods, namely SDEdit [36], P2P [37], and UNSB [38].

The comparison results, as shown in Table 1, indicate that our proposed algorithm achieved the best quantitative evaluation metrics for image generation across all three datasets, particularly demonstrating a significant advantage when applied to the newly developed exhibition design dataset, ExpoArchive. The comparative results underscore the positive contributions of the multimodal encoder module, attention mechanism module, and the fine-tuning of the diffusion model in the style transfer image generation process. Specifically, the multimodal encoder module efficiently processes and integrates data from various modalities—text information, source images, and style images—allowing for precise extraction and representation tailored to the unique characteristics of each modality. Additionally, the attention mechanism, composed of multiple Hierarchical Attention Modules (HAM), adeptly identifies and emphasizes key style and content elements within the images. By feeding the content and style features, enhanced by the attention mechanism, into the U-Net architecture at each step of the diffusion model, this fine-tuning approach ensures precise adjustments of style and content across different regions, thereby guaranteeing that the generated images align with the target style while faithfully representing the content of the source image.

Table 1: Quantitative comparison of the proposed method with other state-of-the-art methods

Methods	ExpoArchive		Summer2Winter		Label2Cityscape		Map2Satellite	
	FID	KID	FID	KID	FID	KID	FID	KID
NOT	289.3	10.28	185.5	8.732	221.3	19.76	224.9	16.59
CycleGAN	157.1	4.32	84.9	1.022	76.3	3.532	54.6	3.43
MUNIT	200.8	7.38	115.4	4.901	91.4	6.401	181.7	12.03
Distance	165.8	6.51	97.2	2.843	81.8	4.41	98.1	5.789
GcGAN	167.9	7.21	97.5	2.755	105.2	6.824	79.4	5.153
CUT	150.5	6.01	84.3	1.207	56.4	1.611	56.1	3.301
SDEdit	174.3	5.28	118.6	3.218	–	–	–	–
P2P	120.5	4.31	99.1	2.626	–	–	–	–
UNSB	110.8	2.55	73.9	0.421	53.2	1.191	47.6	2.013
Ours	87.9	1.98	70.1	0.38	55.1	1.205	45.5	1.898

To highlight the advantages of STConv, we compare it with several state-of-the-art dynamic convolution methods, including AKConv and DyConv. As shown in Table 2, STConv achieves a better balance between computational efficiency and feature extraction capability. Specifically, STConv reduces the FLOPs by 15% compared to AKConv while maintaining a lower FID score on the ExpoArchive dataset.

Table 2: Comparisons between STConv and AKConv

Method	Params (M)	FLOPs (G)	FID (ExpoArchive)
AKConv	12.5	5.8	23.4
DyConv	11.8	5.5	22.8
STConv (Ours)	10.2	4.9	21.5

We further compare our method with state-of-the-art diffusion-based approaches, including ControlNet and InstructPix2Pix. As shown in Table 3, our method achieves superior performance in terms of FID and KID scores, while maintaining lower computational costs.

Table 3: Comparison with diffusion-based methods

Method	FID (ExpoArchive)	KID (ExpoArchive)	FLOPs (G)
ControlNet	24.3	0.012	8.2
InstructPix2Pix	23.8	0.011	7.9
Ours	21.5	0.009	4.9

4.5 Ablation Study

In the ablation study, we utilized CLIP-guided diffusion [39] as the backbone and sequentially integrated the three modules of our proposed algorithm: Multi-modal Decoder (MMD), Attention Module (AM), and Fine-tune Diffusion Module (FDM). We evaluated the quantitative metrics, Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), on the ExpoArchive dataset, with the results presented in Table 4.

Table 4: Quantitative comparison results of ablation experiments on the ExpoArchive dataset

Method	FID	KID
Backbone	197.37	9.36
Backbone + MMD + AM	99.72	2.14
Backbone + MMD + FDM	103.24	3.28
Backbone + AM + FDM	100.19	2.91
Backbone + MMD + AM + FDM	87.9	1.98

The baseline performance exhibited relatively poor results, with elevated FID and KID values. This indicates that utilizing only CLIP-guided diffusion for image generation leads to a significant gap between the generated images and the target style and content, suggesting that the quality of the generated images needs improvement.

The combination of Baseline + MMD + AM resulted in a substantial reduction in both FID and KID metrics. The MMD module effectively processes and integrates data from diverse modalities, providing a richer informational foundation for image generation. Meanwhile, the AM module enhances image quality by simulating the functioning of the human visual system, focusing on crucial areas within the image. The synergistic effect of these two modules allows the algorithm to better merge the target style while preserving the image content.

When combining Baseline + MMD + FDM, the FID metric showed a decrease; however, the KID metric increased compared to the Baseline + MMD + AM configuration. This is attributed to the FDM module's fine-tuning of the diffusion model, which, in the absence of the AM module, lacks precise control over the style and content of different regions, resulting in a certain deviation in style from the target image.

For the combination of Baseline + AM + FDM, both FID and KID metrics decreased, albeit not as significantly as in the Baseline + MMD + AM combination. This indicates that while the AM module plays

a crucial role in capturing key elements of the image, the performance of the FDM module may be limited in the absence of the rich informational foundation provided by the MMD module.

Ultimately, the combination of Baseline + MMD + AM + FDM achieved the lowest values for both FID and KID metrics. This outcome underscores the significant enhancement in algorithm performance facilitated by the collaboration of the three modules. The MMD module offers a wealth of information for image generation, the AM module improves image quality by focusing on critical areas, and the FDM module optimizes the style and content of the generated images through fine-tuning of the diffusion model.

To evaluate the contribution of the semantic encoder, we conduct an ablation study by removing or replacing it with a standard Transformer encoder. As shown in Table 5, the semantic encoder improves the FID score by 2.3 points, demonstrating its effectiveness in fusing spatial and semantic information.

Table 5: Ablation study on semantic encoder

Model variant	FID (ExpoArchive)
Full model	21.5
Without semantic encoder	23.8
Replace with standard transformer	22.7

We compare three fusion strategies: concatenation, additive fusion, and our proposed cross-modal attention. As shown in Table 6, cross-modal attention improves FID by 1.8 points compared to concatenation, demonstrating its effectiveness in capturing nuanced interactions between modalities.

Table 6: Ablation study on multimodal fusion

Fusion strategy	FID (ExpoArchive)	KID (ExpoArchive)
Concatenation	89.7	2.1
Additive fusion	88.5	1.9
Cross-modal attention	86.7	1.7

We evaluate STConv on DTD (textures) and COCO (structural diversity). As shown in Table 7, STConv achieves FID scores of 24.1 (DTD) and 28.3 (COCO), outperforming AKConv by 12% and 9%, respectively. This validates its adaptability to diverse features.

Table 7: Generalization analysis of STConv

Dataset	Method	FID	Texture consistency
DTD	AKConv	27.4	3.1
DTD	STConv	24.1	4.2
COCO	AKConv	31.2	3.8
COCO	STConv	28.3	4.5

4.6 Qualitative Results

To visually demonstrate the performance of the proposed model in the context of generating images for exhibition design style transfer tasks, we selected a series of representative images for qualitative assessment. These images encompass various themes, styles, and complexities, aimed at comprehensively evaluating the model's generalization capabilities and adaptability.

The results illustrate the model's effectiveness in integrating the target style while maintaining the integrity of the image content. As shown in Fig. 10, the key elements of the original images remain distinctly recognizable following style transfer, while the textures, colors, and brushstroke characteristics of the target style are seamlessly incorporated. This harmonious unification of style and content highlights the model's robust ability to capture and fuse image features effectively.

Figure 10: Illustration of the style transfer image generation effects of the proposed algorithm applied to architectural and natural landscape images. **The first column** displays the source images, **the second column** shows the images transferred to the target style, and **the thirdx, fourth,** and **fifth columns** illustrate the transfer of textual content (specific details are provided in the captions)

Furthermore, the model's performance in processing images related to exhibition design is showcased. In Fig. 11, the original images feature multiple layers and intricate details. During the style transfer, the model not only preserves the clarity and contrast of these layers but also skillfully incorporates unique elements of the target style, resulting in images that retain the original structural information while exuding a new artistic flair. A comparison between the original and style-transferred images clearly reveals the model's adjustments in color, brushwork, and texture, ensuring that the final generated images maintain the contours and characteristics of the source images in exhibition design, while also conveying the depth and richness intended by the style image and accompanying text.

Figure 11: Illustration of the style transfer image generation effects of the proposed algorithm applied to images related to exhibition design. **The first column** presents the source images, **the second column** demonstrates the images transferred to the target style, and **the third, fourth,** and **fifth columns** depict the transfer of textual content (specific details are provided in the captions)

4.7 Quantitative Analysis of Style Transfer Details

To evaluate the quality of style transfer, we conduct a user study with 50 participants and calculate the perceptual loss. As shown in Table 8, our method achieves higher scores in both texture consistency and content fidelity compared to baseline methods.

Table 8: Quantitative analysis of style transfer

Method	Texture consistency	Content fidelity	Perceptual loss
CycleGAN	3.2	3.5	0.45
CUT	3.8	3.7	0.38
Ours	4.5	4.3	0.28

4.8 Applications and Limitations of the Proposed Method

While our method achieves promising results in single-image style transfer, its applicability in 3D scenes and dynamic interactions remains to be explored. Challenges include maintaining multi-view consistency and meeting real-time processing requirements. However, its current limitations in handling 3D scenes and dynamic interactions highlight the need for further research. These challenges also present opportunities

for future advancements in exhibition design and related fields. Future work will focus on extending the algorithm to handle 3D data and optimizing its computational efficiency for real-time applications.

4.9 Mixed-Style Transfer Analysis

To evaluate the effectiveness of the Hybrid Attention Module (HAM) in handling mixed-style scenarios, we conducted experiments on a mixed-style dataset combining 50% classical and 50% modern styles. The results demonstrate the superiority of our approach in terms of both quantitative metrics and qualitative assessments.

As shown in Table 9, our method achieves the lowest FID score (24.5), indicating better alignment with the target style distribution compared to CycleGAN (34.2) and StyTr2 (29.8). The HAM module achieves a style consistency score of 4.3/5, outperforming CycleGAN (3.1) and StyTr2 (3.8), demonstrating its ability to accurately capture and transfer mixed styles. With a content fidelity score of 4.5/5, our method preserves the structural integrity of the source image better than the baseline methods.

Table 9: Performance comparison on mixed-style transfer

Method	FID (Mixed-Style) ↓	Style Consistency (1–5) ↑	Content Fidelity (1–5) ↑
CycleGAN	34.2	3.1	3.4
StyTr2	29.8	3.8	3.9
Ours (HAM)	24.5	4.3	4.5

4.10 Fine-Grained Evaluation

To holistically evaluate the generated images, we introduce three additional metrics: Learned Perceptual Image Patch Similarity (LPIPS), which measures perceptual similarity between generated and real images at the patch level; mIoU which evaluates semantic consistency by comparing segmentation masks of source and stylized images using a pre-trained DeepLabV3 model; User Scores, five participants rate style consistency, content preservation, and aesthetic quality on a 1–5 scale (higher score indicates better). And the comparison between the proposed method and CycleGAN, StyTr2 is shown as Table 10. Our method achieves the lowest LPIPS score (0.28), indicating superior perceptual quality. The highest mIoU (78.5%) of our method also confirms strong semantic consistency. Moreover, User Score rates our method highest (4.5/5) in aesthetic quality.

Table 10: Fine-grained evaluation

Method	LPIPS	mIoU	User score
CycleGAN	0.45	65.2	3.2
StyTr2	0.38	72.4	3.9
Ours	0.28	78.5	4.5

To address the diversity of generated designs for exhibition halls, we introduce two new quantitative metrics: Intra-Style Diversity (ISD) and Inter-Style Diversity (ITD). ISD measures the variation among images generated from the same style input, while ITD quantifies differences between outputs from different style inputs. The comparison results are shown in Table 11. When transferring a “Futuristic” style to 50 distinct

exhibition hall layouts, our method achieves an ISD of 0.38 (higher than CycleGAN's 0.12 and StyTr2's 0.25), demonstrating its ability to produce diverse designs under a single style constraint.

Table 11: Diversity comparison on ExpoArchive dataset

Method	ISD	ITD
CycleGAN	0.12	0.28
StyTr2	0.25	0.41
Ours	0.38	0.67

5 Conclusion

This paper presents a semantic-enhanced multimodal style transfer algorithm specifically designed to address the diverse stylistic and visual consistency needs of exhibition hall design. By employing a multimodal encoder architecture, this approach effectively extracts and integrates features from text, source images, and style images, providing a comprehensive foundation for style transfer. The proposed STConv convolutional kernel and Transformer encoder enhancements allow the algorithm to capture various style and content features with flexibility and precision. Additionally, a hybrid attention module accurately aligns content and style features, ensuring the integrity and visual harmony of the transferred images. Experimental results show that the proposed method outperforms traditional style transfer techniques in terms of visual quality, stylistic coherence, and aesthetic appeal of generated images. Tests on the ExpoArchive dataset, among others, highlight the algorithm's ability to handle different styles, complex structures, and multi-level semantic features. The significant improvements in FID and KID scores further validate the effectiveness of the algorithm in generating high-quality style transfer images.

Future research will focus on refining the algorithm's control over fine-grained stylistic and semantic features, as well as enhancing computational efficiency for real-time exhibition design and interactive applications. Expanding the diversity of exhibition design datasets and enriching annotation data will also aid in improving the algorithm's generalization and practical value.

The algorithm proposed in this study has significant practical application value in the field of exhibition hall design, providing high-quality style transfer images for such designs. Despite achieving notable results, there are still some limitations, such as the need to improve computational efficiency and further enhance the control over fine-grained style features.

Acknowledgement: We extend our heartfelt gratitude to Dr. Yuting Wang from the School of Software at Northeastern University for her invaluable support and guidance throughout our research. We would also like to acknowledge the Cross-Media Artificial Intelligence Laboratory for their exceptional technical assistance and provision of hardware services, which were critical to the successful completion of this project. Their contributions were instrumental in advancing our research and achieving our goals.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Qing Xie; data collection: Qing Xie; analysis and interpretation of results: Qing Xie; draft manuscript preparation: Ruiyun Yu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 2414–23. doi:10.1109/CVPR.2016.265.
2. Yeh MC, Tang S, Bhattad A, Forsyth DA, Yeh MC, Tang S, et al. Quantitative evaluation of style transfer. arXiv:1804.00118. 2018.
3. Ma L, Li N, Zhu P, Tang K, Khan A, Wang F, et al. A novel fuzzy neural network architecture search framework for defect recognition with uncertainties. IEEE Trans Fuzzy Syst. 2024;32(5):3274–85. doi:10.1109/TFUZZ.2024.3373792.
4. Liu S, Lin T, He D, Li F, Wang M, Li X, et al. AdaAttN: revisit attention mechanism in arbitrary neural style transfer. arXiv:2108.03647. 2021.
5. Luo X, Han Z, Yang L, Zhang L. Consistent style transfer. arXiv:2201.02233. 2022.
6. Efros AA, Leung TK. Texture synthesis by non-parametric sampling. In: Proceedings of the 7th IEEE International Conference on Computer Vision; 1999 Sep 20–27; Kerkyra, Greece. p. 1033–8. doi:10.1109/ICCV.1999.790383.
7. Gooch B, Gooch A. Non-photorealistic rendering. Boca Raton, FL, USA: AK Peters/CRC Press; 2001.
8. Elad M, Milanfar P. Style transfer via texture synthesis. IEEE Trans Image Process. 2017;26(5):2338–51. doi:10.1109/TIP.2017.2678168.
9. Mordvintsev A, Olah C, Tyka M. Inceptionism: going deeper into neural networks. [cited 2025 Jan 1]. Available from: https://research.google/blog/inceptionism-going-deeper-into-neural-networks/.
10. Li C, Wand M. Combining Markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 2479–86. doi:10.1109/CVPR.2016.272.
11. Radford A, Metz L, Chintala S, Dinakaran R, Easom P, Zhang L, et al. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:1511.06434. 2015.
12. Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Proceedings of the International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia. p. 214–23.
13. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 4396–405. doi:10.1109/cvpr.2019.00453.
14. Yao Y, Ren J, Xie X, Liu W, Liu YJ, Wang J. Attention-aware multi-stroke style transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 1467–75. doi:10.1109/CVPR.2019.00156.
15. Deng Y, Tang F, Dong W, Ma C, Pan X, Wang L, et al. StyTr2: image style transfer with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 11316–26. doi:10.1109/CVPR52688.2022.01104.
16. Delbracio M, Milanfar P. Inversion by direct iteration: an alternative to denoising diffusion for image restoration. arXiv:2303.11435. 2023.
17. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst (NeurIPS). 2020;33:6840–51.
18. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B, Mantri KSI, et al. High-resolution image synthesis with latent diffusion models. arXiv:2112.10752. 2021.
19. Zhang X, Song Y, Song T, Yang D, Ye Y, Zhou J, et al. AKConv: convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters. arXiv:2311.11587. 2023.
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv:1706.03762. 2017.

21. Wan Q, Huang Z, Lu J, Yu G, Zhang L. SeaFormer: squeeze-enhanced axial transformer for mobile semantic segmentation. In: International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda.
22. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: IEEE International Conference on Computer Vision; 2021 Oct 10–17; Montreal, QC, Canada.
23. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W. CCNet: crisscross attention for semantic segmentation. In: IEEE International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea.
24. Ho J, Kalchbrenner N, Weissenborn D, Salimans T. Axial attention in multidimensional transformers. arXiv:1912.12180. 2019.
25. Summer2Winter Yosemite: CycleGAN's Summer Winter Images Dataset at Yosemite. [cited 2025 Jan 1]. Available from: https://www.kaggle.com/datasets/balraj98/summer2winter-yosemite.
26. The Cityscapes Dataset: 5 000 images with high quality annotations •20 000 images with coarse annotations •50 different cities. [cited 2025 Jan 1]. Available from: https://www.cityscapes-dataset.com/.
27. Larger Google Sat2Map dataset. [cited 2025 Jan 1]. Available from: https://github.com/taesungp/larger-google-sat2maps-dataset.
28. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. arXiv:1706.08500. 2017.
29. Chen R, Huang W, Huang B, Sun F, Fang B. Reusing discriminators for encoding: towards unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 8165–74. doi:10.1109/CVPR42600.2020.00819.
30. Korotin A, Selikhanovych D, Burnaev E. Neural optimal transport. In: International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda.
31. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. doi: 10.1109/ICCV.2017.244.
32. Huang X, Liu MY, Belongie S, Kautz J. Multimodal unsupervised image-to-image translation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany.
33. Benaim S, Wolf L. One-sided unsupervised domain mapping. arXiv:1706.00826. 2017.
34. Fu H, Gong M, Wang C, Batmanghelich K, Zhang K, Tao D. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 2422–31. doi:10.1109/cvpr.2019.00253.
35. Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for unpaired image-to-image translation. In: European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. p. 319–45.
36. Meng C, He Y, Song Y, Song J, Wu J, Zhu JY, et al. SDEdit: guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations; 2022 Apr 25–29; Virtual.
37. Hertz A, Mokady R, Tenenbaum J, Aberman K, Pritch Y, Cohen-Or D. Prompt-to-prompt image editing with cross attention control. arXiv:2208.01626. 2022.
38. Kim B, Kwon G, Kim K, Ye JC. Unpaired image-to-image translation via neural schrödinger bridge. In: International Conference on Learning Representations; 2024 May 7–11; Vienna, Austria.
39. Katherine C. Clip-guided diffusion. [cited 2025 Jan 1]. Available from: https://github.com/afiaka87/clip-guided-diffusion.