Cross-Modal Attention GAN for Text-to-Artistic Image Generation

Huibin Jin
Department of Architecture, Henan Technical College of Construction, Zhengzhou 450064, China
E-mail: jinhuibin1983@163.com

Keywords: cross-modal, attention, textual description, art and design images, channel attention mechanism, image refinement

Received: February 17, 2024

Text and image are two very different data modalities. In the process of converting text-to-image, the essential difference between the two modalities leads to low R-value accuracy (a metric for semantic consistency) in generating the final image. In order to improve the relevance and artistry of the generated design images, a cross-modal attention-based method for generating artistic design images from text is investigated. A cross-modal attention-based generative adversarial network model (CMAGAN) is constructed to realize text-generated art and design images. The CMAGAN model is divided into three phases: in the initial cross-modal image generation phase, the pre-trained RNN is used to encode text descriptions, obtain sentence and word feature vectors, and generate initial cross-modal art and design images through the content-aware up-sampling module and channel-attention convolution module; In the initial image refinement stage, spatial and channel attention mechanisms are introduced to accurately match text and image features; in the image refinement stage, the image features are further attended to and fused using the secondary attention (AoA) mechanism to enhance the visual features; and under the effect of the integrated loss function, the semantically consistent, high-quality and richly-detailed art design images are obtained from the textual descriptions. Conduct experiments using datasets consisting of 6000 images each from Artstor and DeviantArt, two online art platforms. The ablation experiment showed that the complete model had the highest R-value accuracy (such as 0.86 for art, furniture, advertising, and graphics), the highest initial score (IS) (such as 3.65 for art), and the lowest Frechet Inception distance score (FID) (such as 20.4 for art) when generating art, furniture, advertising, and graphics design images. This indicates that the generated images have the strongest semantic consistency with the input text, are clearer and more diverse, and have the shortest distribution distance from real images. Compared with existing methods, the proposed method exhibits better scores in visual semantic similarity (VSS) across multiple sample sizes, with a more significant improvement and consistently maintaining a high level. The above results fully verify the advantages of our method in generating artistic design images from text.

Povzetek: Opisan je CMAGAN, GAN-model s čezmodalno pozornostjo za generiranje umetniških slik iz besedila, ki dosega visoko semantično skladnost, vizualno kakovost in podrobnost brez dodatnega treniranja.

1 Introduction

In today's era of deep integration between art and technology, this technology not only represents a revolution in the way art is created, but also quietly changes the way we acquire and experience art. For the general public, it means being able to stimulate creativity more intuitively through written descriptions, instantly transforming their imagination into visual works of art, greatly enriching the possibilities for personal expression and creation. The traditional hand-drawn or software drawing method can create exquisite works of art, often requiring a significant amount of time and energy. The text to generate art design image technology, according to the user's text description, automatically generate images to meet the requirements, greatly improving the creative efficiency. The technology can also be customized according to the user's personalized needs, to meet the user's pursuit of uniqueness and differences. In addition, in industries such as advertising, media, and gaming, the efficient application of this technology can bring more personalized and precise visual content, enhance the attractiveness and interactivity of information dissemination, and enable non professional users to enjoy the convenience and fun brought by technological innovation, thus making it easier to integrate creative elements into daily life and work, improving the quality of life and work efficiency [1-2].

However, it is not easy to realize the technique of generating art and design images from text. Currently, some methods have been proposed and applied to this field. For example, Endo [3] achieved text-to-art-design image generation control without additional training by manipulating the cross-attention graph of a pre-trained diffusion model. The method introduces a mask attention guidance strategy, which indirectly controls the attention of each word and pixel and adjusts the noisy images input into the diffusion model, thereby generating artistic design images that better align with the semantic masks. Although text to art design image generation control has been achieved, it requires manipulating the pre trained model's cross attention map, which increases implementation complexity. Bahani et al [4] designed three different architectures to compare the performances of T5, GPT-2, and BERT in the task of text-to-image (T2I) generation. These pre-trained transformer models were first fine-tuned to generate text vectors corresponding to the input text descriptions. These text vectors are then passed as inputs to the generator of DF-GAN (a deep feature generation adversarial network) using affine transformations to convert them into images. However, this method lacks cross modal information fusion, resulting in insufficient semantic consistency and detail representation of the generated images. Nezhad et al [5] by combining Generative Adversarial Network (GAN), deep learning techniques and user opinions, first evaluates the popularity of the image itself by using a popularity module, and then The unpopular images are converted into popular ones by a conversion module. In this process, the model can also perform multidimensional and multidomain image conversion, and finally generate art and design images that meet the user's aesthetics with high quality and diversity. However, this method relies on user feedback, which limits the innovation and uniqueness of the images. Watanabe et al. [6] realized the automatic generation of art design images from creative text by applying the visual language model's ability to learn small samples in context and focusing on the accurate matching and diversity of image features and text descriptions. It can not only capture the artistic ideas and style requirements in the text, but also integrate the visual elements in the images to create creative art and design images that conform to the textual descriptions. This method relies on small sample learning, and insufficient samples can affect the quality and consistency of the generated images. The current research progress is shown in Table 1.

Table 1: Comparison of research progress

Research method	Method Description	Limitation	How can this method overcome these limitations
Endo [3]	By manipulating the cross-attention map of the pre trained diffusion model, the generation control of text to art design images is achieved, and a masked attention guidance strategy is proposed	Need to have a deep understanding of the internal working principles of the model, take additional steps to adjust the attention map, and increase the complexity of the model implementation	This article constructs a Generative Adversarial Network model based on cross modal attention, which does not require manipulating the internal attention map of the pre trained model. The image is gradually generated through three stages, reducing implementation complexity
Bahani et al. [4]	Design three different architectures to compare the performance of T5, GPT-2, and BERT in T2I generation tasks, fine tune the pre trained transformer model to generate text vectors, and use affine transformation to pass to DF-GAN to generate images	Lack of cross modal information fusion leads to a lack of semantic consistency or detail representation in the generated images	This article proposes a cross modal attention mechanism to accurately match text and image features at different stages of image generation, enhancing semantic consistency and detail representation
Nezhad et al. [5]	Combining GAN, deep learning techniques, and user feedback, evaluate the popularity of images, convert unpopular images into popular images, and perform multidimensional and multi domain image conversion	Relying on user feedback and limiting image innovation and uniqueness may result in a lack of novelty and uniqueness	This method does not rely on user opinions and automatically learns the association between text and images through a cross modal attention mechanism, generating innovative and unique art and design images
Watanabe et al. [6]	Utilizing the small sample learning ability of visual language models in context to achieve automatic generation of creative text to art design images, capturing artistic concepts and style requirements in the text	Small sample learning in context dependency requires a certain number of relevant images and text pairs for training or adaptation. Insufficient or diverse samples can affect the quality and consistency of generated images	This method does not rely on small sample learning. By constructing a generative adversarial network model based on cross modal attention, under the comprehensive loss function, it generates art design images that are semantically consistent, of high quality, and rich in details with text descriptions

In order to overcome the limitations of the above methods, the cross-modal attention mechanism has been introduced into the text generation art design image technology. Cross-modal attention mechanism is a technology that can simultaneously process two modal information of text and image, which can focus on the key information in the text in the process of generating images and correlate it with the corresponding region in the image [7]. Through this mechanism, we can generate images that are more consistent with the text description and more artistic. Therefore, in order to solve the problem of low R-value accuracy in generated images caused by the essential differences between text and image modalities, this paper proposes a cross-modal attention-based method for generating artistic design images from text. By constructing a generative adversarial network model based on cross modal attention, there is no need to manipulate the internal attention map of the pre trained model. The image is gradually generated through three stages, reducing complexity; At the same time, the cross modal attention mechanism accurately matches text and image features, enhancing semantic consistency and detail representation; In addition, the method proposed in this article does not rely on user opinions or small sample learning, and automatically learns the association between text and images, generating high-quality art and design images that are innovative, unique, and consistent with text descriptions.

2 Text Generating art and design image methods

2.1 A generative adversarial network model based on cross-modal attention

A cross-modal attention-based generative adversarial network model (CMAGAN) is designed to make full use of the mechanism of cross-modal attention to achieve accurate generation from text to art design images. The model is mainly divided into three stages: image cross-modal initial generation stage, image initial refinement stage and image re-refinement stage. The results of the model network are shown in Figure 1.

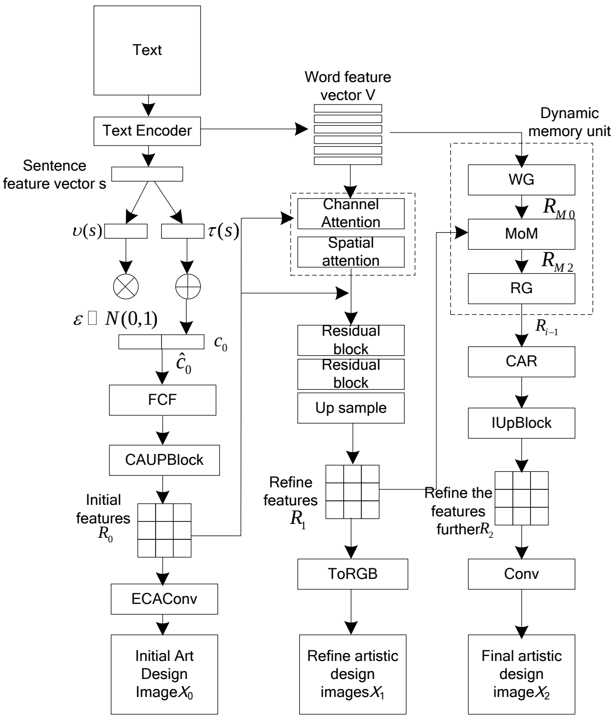


Figure 1: Structure of generative adversarial network model (CMAGAN) based on cross modal attention

The components in Figure 1 constitute multiple key modules in the image generation process, which work together to optimize the entire process from initial feature extraction to final image generation, especially playing an important role in the CMAGAN model. Specifically, CAUPBlock combines channel attention and upsampling operations to extract and optimize image features in the initial generation stage [8]; RMO (Residual Module) alleviates the gradient vanishing problem in deep networks through residual connections, helping the model better learn and transmit feature information; MoM (Multi level Attention Mechanism) captures feature information at different levels, enhances the expressive power of the model, and ensures that the generated images are rich in details and semantically consistent; CAR (Channel Attention Residual) combines channel attention and residual connections to further refine features, ensuring high-quality and semantic consistency of generated images; ECAConv, as an efficient channel attention convolution operation, enhances important channel information and improves feature expression ability during feature extraction; FCF (Feature Fusion Block) fuses features from different levels or sources to generate richer feature representations; The ToRGB module converts the feature map into an RGB image and outputs the final image; WG (Weighted Gate) dynamically adjusts the importance of different features through a weighted gating mechanism to optimize the feature fusion process; IUpBlock combines upsampling and feature optimization to gradually improve the resolution and details of the image; Conv, as a standard convolution operation, is responsible for extracting and transmitting feature information.

In the initial generation stage of cross modal images, the model encodes text descriptions through pre trained recurrent neural networks (RNNs) to obtain sentence feature vectors and word feature vectors, and generates initial art design images using content aware upsampling modules and channel attention convolution modules. Entering the initial refinement stage of the image, the model introduces spatial attention and channel attention mechanisms to accurately match text descriptions with artistic design image features. In the image refinement stage, the model utilizes the Attention on Attention (AoA) mechanism to further pay attention to and fuse the image features from the initial refinement stage, enhancing visual features. In addition, the CMAGAN model comprehensively considers multiple aspects such as generation loss, conditional enhancement loss, and DAMSM loss in the loss function to ensure that the generated art design image is semantically consistent with the text description, while improving the realism and artistry of the art design image. These components and mechanisms enhance the realism, diversity, and semantic consistency of images through attention mechanisms, residual connections, and feature fusion.

2.2 Image cross-modal initial generation phase

In the initial generation phase of the art design image, the given text description is fed into the text encoder to get the sentence feature vector $\boldsymbol{s}$ and word vectors $\boldsymbol{V}$, the text encoder used is a pre-trained Recurrent Neural Network (RNN). The sentence feature vectors $\boldsymbol{s}$ is a vector containing semantic features of text sentences for cross-modal initial generation of art and design images. The word vector $\boldsymbol{V}$ is a vector containing semantics of 18 words, which is used for image generation in the initial refinement and re-refinement stages of the image [9]. For the sentence feature vector $\boldsymbol{s}$ obtained by encoding, conditional enhancement is required, which is to obtain the mean matrix $\tau(\boldsymbol{s})$ and the covariance matrix $\upsilon(\boldsymbol{s})$ from the Gaussian distribution $\square\left[\tau(\boldsymbol{s}),\sum(\boldsymbol{s})\right]$ of the sentence eigenvector $\boldsymbol{s}$, then compute to obtain the eigenvectors $\boldsymbol{c}_0$, ($\boldsymbol{c}_0 = \tau(\boldsymbol{s}) + \upsilon(\boldsymbol{s}) \square \varepsilon$, $\square$ represents a dot product operation, $\varepsilon \square N(0,1)$), and finally $\boldsymbol{c}_0$ and a random noise Z sampled from a normal distribution spliced together to obtain $\hat{\boldsymbol{c}}_0$. After $\hat{\boldsymbol{c}}_0$ performs a full connection operation, the content sensing upsampling module is input, and the feature map R_0 is obtained after upsampling, the cross-modal initially generated art design image is obtained after the feature map is fed into the channel attention convolution module.

2.2.1 Content-aware upsampling module

Before the initial generation of art design images, feature maps need to be up-sampled, and the current common up-sampling methods include nearest neighbor interpolation and inverse convolution. However, the sense field of nearest neighbor interpolation is too small and does not utilize the semantic information, while the inverse convolution is too large in computation. In this paper, the content-aware up-sampling module uses the original feature map to obtain the reorganized convolution kernel, and uses the reorganized convolution kernel to up-sample the input feature map, which takes into account the relationship between each pixel and the surrounding area, and at the same time avoids the problem of too many parameters and too much computational effort [10]. The content-aware up-sampling module consists of an adaptive convolutional kernel prediction module and a content-aware feature reorganization module. After the feature map is input into the content-aware up-sampling module, the up-sampling operation is repeated for four times, assuming that the dimension of the input feature map R is $C \times W \times H$, with the upsampling multiplier set to S (set to 2 in this paper). Among them, choosing to repeat upsampling four times is because it can significantly improve the resolution of the image without introducing excessive computational burden, making the generated image more delicate and realistic. And setting the upsampling multiplier to 2 is because this value can effectively double the size of the feature map during each upsampling, so that after four upsampling, the size of the feature map can be increased to 16 times the original size. This is an efficient and practical multiple selection for generating high-resolution images from low resolution feature maps, which ensures a significant improvement in image quality and controls the consumption of computational resources. After the content-aware up-sampling module outputs the new feature map R' after up-sampling, whose dimension is $C \times SW \times SH$, the reign $l' = (i', j')$ of output feature maps R', in which corresponds to $l = (i, j)$ in the input feature map $R = f_{\text{conv}n}\left\{f_{\text{conv}n-1}\left[\cdots f_{\text{conv}1}(\boldsymbol{V}, \boldsymbol{s})\cdots\right]\right\}$, the correspondence is $i = \left[i'/S\right], j = \left[j'/S\right]$. Among them, $f_{\text{conv}n}$ represents the convolution operation of the *n*-th layer.

After inputting the feature figure R, the convolution kernel $\gamma_{l'}$ is predicted from each region l' of the output feature figure R' in the adaptive convolution kernel prediction module ψ. As shown in Equation (1), the original feature diagram is multiplied in the content aware feature reorganization module ξ and the predicted convolution kernel to obtain the result, as shown in Equation (2):

$$\gamma_{l'} = \psi\left[Z\left(R_l, k_{encoder}\right)\right] \tag{1}$$

$$R'_{l'} = \xi\left[Z\left(R_l, k_{up}\right), \gamma_{l'}\right] \tag{2}$$

Where, $Z\left(R_l, k_{up}\right)$ represents a subregion of size $k_{up} \times k_{up}$ around the midpoint l of the feature graph R. $k_{encoder}$ indicates the size of the content encoder.

In the adaptive convolutional kernel prediction module, the feature map first undergoes a 1×1 convolutional layer to reduce the number of channels from C to C_m, and then the convolutional kernel is predicted by a content encoder with the number of input channels as C_m, and the number of output channels as $C_{up} = S^2 k_{up}^2$, expanding the channel dimension in the space dimension, we obtain a recombined convolutional kernel with the size of $SH \times SW \times k_{up}^2$, and finally normalized using the softmax function, so that the weights of the recombined convolutional kernel sum to 1.

For each position l' in the output feature map, the content-aware feature reorganization module maps it back to the input feature map to take out the region of size of $k_{up} \times k_{up}$ centered on $l = (i, j)$, and the recombination convolution kernel predicted at that point is made as a dot product to obtain the output value, as shown in Equation (3), where different channels at the same location share the same recombination convolution kernel.

$$R'_{l'} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} \gamma_{l'(n,m)} \cdot R_{(i+n, j+m)} \quad (3)$$

Of which: $l = (i, j)$ is the point of the output feature map at the corresponding position on the input feature map; $r = \left\lfloor k_{up} / 2 \right\rfloor$ is a neighborhood of l.

The content-aware upsampling module, composed of an adaptive convolution kernel prediction module and a content-aware feature reorganization module, addresses the issues of small receptive fields and lack of semantic information utilization in nearest-neighbor interpolation, as well as the high computational cost of deconvolution. After the feature map is input, the upsampling operation is repeated four times with an upsampling rate set to 2, thereby enhancing image resolution while controlling computational resource consumption. In the adaptive convolution kernel prediction module, the feature map undergoes channel compression through a convolutional layer, and the reorganization kernels are predicted by a content encoder, followed by normalization. The content-aware feature reorganization module then utilizes these reorganization kernels to perform mapping and dot product operations for each position of the input feature map, generating the output values. By considering the relationship between each pixel and its surrounding regions, the entire module achieves efficient and practical upsampling, producing a new feature map with increased dimensions and improved quality.

2.2.2 Channel attention convolution module

After upsampling, the feature map is obtained and input into the generator, which generates artistic design images across modalities through convolution operations. By using channel attention to weight feature maps, the generated art and design images are enriched in details [11]. The channel attention convolution module is a key component used for cross modal generation of art design images after upsampling the feature map and inputting it into the generator. This module weights the feature maps through channel attention mechanism to enhance the detail richness of the generated images. The calculation of channel attention weights is based on the result of global average pooling of the input feature map. Global average pooling calculates the average pixel value of each channel's spatial position, compressing the feature map into a single numerical value. The size of the weight matrix is, and for each channel, its weight calculation only considers the adjacent 5 channels. This mechanism not only enriches the details of the image, but also maintains performance while reducing model complexity through cross channel interaction.

In the channel attention convolution module, the channel attention weights ω is calculated as shown in Equation (4):

$$\omega = \sigma(Qy) \quad (4)$$

Of which: $y = G_{GAP}(R)$, obtained from the input feature map R after global average pooling [12]; Q is the weight; σ is then the Sigmoid function. Global average pooling is a technique for dimensionality reduction of feature maps. For each channel of the feature map, the average pixel value of all spatial positions (i.e., width and height directions) is calculated to compress the feature map of that channel into a single numerical value. Assuming that the accepted feature map $R \in \square^{W\times H\times C}$, W, H and C represent the width, height and channel dimensions of the feature map, respectively. The global average pooling Equation is shown in Equation (5) as follows:

$$G_{GAP}(R) = \frac{1}{WH} \sum_{i=1, j=1}^{W,H} R_{ij} \quad (5)$$

The size of weighting Q is $k \times C$, for each channel y_i, corresponding to the calculation of weights ω_i only needs to take into account the neighboring k channels (set to 5 in this paper), as shown in Equation (6):

$$\omega_i = \sigma(\sum_{j=1}^{k} w^j y_i^j), y_i^j \in \Omega_i^k \quad (6)$$

Where, Ω_i^k represents the set of k channels adjacent to the channels y_i; w^j is the weight of the j adjacent channels y_i^j of y_i.

2.3 Preliminary image refinement stage

In the initial image refinement stage, spatial attention and channel attention [13] are introduced to accurately match words with the features of the cross-modal initially generated image and select the image information most relevant to the textual description, so as to avoid differences in the contribution of different words to the image content affecting the generation of art design images. The spatial attention module accurately focuses on the spatial region that is closely related to the text description by calculating the similarity between the text features and image features, thus enhancing the semantic consistency of the image; the channel attention module dynamically selects the channel features that are the most relevant to the text description to help the network focus more on the channel features that are closely related to the text description to further enhance the consistency between the image and the text [14].

The spatial attention module text word vector $\boldsymbol{V}$ and image features R_0 as input. First, the perceptual layer is used to map word vector $\boldsymbol{V}$ to the common semantic space c of image features. Simultaneously selecting the image feature R_0 used to compute the vectors between different subregions and words as shown in (7) and (8), where:

$$\beta_{j,i} = \frac{\exp(s'_{j,i})}{\sum_{k=0}^{T-1} \exp(s'_{j,k})} \tag{7}$$

$$d_j = \sum_{i=0}^{T-1} \beta_{j,i} c_j \tag{8}$$

Where, $s'_{j,i} = m_j^T c_i$ represents the attentional weight of the j -th hidden feature in the image and the i -th subspace in the text. c_j represents the semantic value noticed by the j -th subregion in the text, and $\beta_{j,i}$ represents the i -th weight value of the j -th subregion of the model on the image feature map. Then, the computed feature map $D = (d_0, d_1, ...d_{n-1})$ is used as input for the next stage.

Two inputs are still accepted in the channel attention phase: word features $\boldsymbol{V}$ and image features $D = (d_0, d_1, ...d_{n-1})$. First, apply the word features $\boldsymbol{V}$ to acquire the spatial features u corresponding to the image features through the perceptual layer, after obtaining the spatial feature u through 1×1 convolution u', then performing the dot product multiplication with the image feature D and maximizing the activation function to obtain the channel attention weight matrix a, finally, the attention weight a of the channel and the semantic spatial feature u' are processed [15] again to obtain the image feature out of the space and the channel content; Again, after residual blocking and upsampling to get the refined image X_1. The main definitions are as follows Equations (9) and (10):

$$u' = con(u) \tag{9}$$

$$a = soft\max(h \times u') \tag{10}$$

$$out = u' \times a \tag{11}$$

Among them, con denotes the 1×1 convolutional block used to obtain semantic value u' corresponding to the image features, a is the attention weight adjusted by the activation function.

2.4 Image refinement stage again

In the stage of fine generation of art and design images, the CMAGAN model utilizes dynamic memory attention units to further process the features of the initial refinement stage of the image, in order to enhance the visual features of the image. This unit selects relevant word features through memory writing and stores them in a dynamically updated feature space. The MoM module is used for two visual feature enhancements, first supplementing local semantic information through self attention mechanism to obtain a first memory, and then fusing features to obtain a second memory. Finally, the response gating selects important image feature information for fusion and outputs brand new response image features. After this stage, art and design images with richer detailed information can be generated.

As shown in Figure 1, the image re-refinement stage in the CMAGAN model utilizes a dynamic memory attention unit to mitigate the effect of the image features generated in the initial image refinement stage on the current generation effect when the visual feature expression is unsatisfactory. The dynamic attention unit utilizes the Writing Gate (WG) for dynamic memory writing, the MoM for visual feature enhancement, and the the response gate (Response Gate, RG) performs a secondary fusion response to process the image feature R_{i-1} generated in the previous stage and feature vector $\boldsymbol{V}$ generated by the encoding [16]. The whole unit processing can be expressed as follows:

$$R_i = RG(MoM(WG(\boldsymbol{V}, R_{i-1}))) \tag{12}$$

In the Equation (12), R_{i-1} denotes the image features generated by the previous generator; R_i represents the image features that have undergone the dynamic memory attention unit to realize visual feature enhancement. The entire dynamic memory attention algorithm is represented as follows.

Input: Word features and image features generated by the previous generator.

Output: Image features R_i after feature enhancement.

Step 1: Global average pooling is performed to the input R_{i-1}, and global average pooled features R_{i-1}^{avg} is obtained.

$$R_{i-1}^{avg} = GAP(R_{i-1}) \tag{13}$$

In the Equation (13), $GAP(\cdot)$ indicates global average pooling.

Step 2: Calculate the importance of each word feature vector $\boldsymbol{V}_i$ to the image feature R_{i-1}, and obtain the importance weight g_i^w. g_i^w as a dynamic memory, write into the gating, to select the words that are associated with the initial refined image to be deposited into the dynamic memory feature space as a dynamic memory R_{M0}. The mathematical expression is:

$$g_i^w = \sigma(A * w_i + B * R_{i-1}^{avg}) \tag{14}$$

$$R_{M0} = U_w(\boldsymbol{V}_i) * g_i^w + U_m(R_{i-1}^{avg}) * (1 - g_i^w) \tag{15}$$

In the Equation (14) and (15), A and B denote the 1×256 and 1×64 matrices, respectively. U_w and U_m denotes a 1 × 1 convolution with different parameters.

Step 3: Read the dynamic memory feature space R_{M0} through the MoM module, and perform attention and fusion to R_{i-1} and R_{M0}, to obtain the image features R_{M2} after secondary visual feature enhancement. The mathematical representation is as follows:

$$R_{M1} = Attn(R_{M0}, R_{i-1}) \tag{16}$$

$$R_{M2} = F([R_{M1}; R_{i-1}]) \tag{17}$$

In the Equation (16) and (17), R_{M1} indicates a memory acquired through an attentional mechanism. $Attn(\cdot)$ indicates an attention operation. $[;]$ denotes the feature splicing operation; $F(\cdot)$ denotes a 1×1 convolution, which is used to fuse R_{M1} and R_{i-1}.

Step 4: Using response gating g_i^r to choose the important image feature information in the R_{M2} and R_{i-1} for fusion [17], and response to output a new response image feature R_i.

$$g_i^r = \sigma(M[R_{M2}; R_{i-1}] + b) \tag{18}$$

$$R_i = R_{M2} * g_i^r + R_{i-1} * (1 - g_i^r) \tag{19}$$

In the Equation (18) and (19), σ denotes the Sigmoid activation function; M denotes the weight in the linear operation; b indicates a bias term.

The MoM module of the dynamic attention unit is mainly through R_{M0} to perform two visual feature enhancements of R_{i-1}. The inputs of R_{M0} and R_{i-1} to MoM were used to supplement the missing local semantic information of R_{i-1} to achieve the first visual feature enhancement and obtain a single memory R_{M1}. And then R_{M1} and R_{i-1} to fuse, realize the second visual feature enhancement, and obtain the second memory R_{M2}, i.e., the image features after two visual feature enhancements.

After the image re-refinement stage, it is possible to obtain art and design image generation results that are richer in detail information.

2.5 Loss function

The loss function of the CMAGAN model consists of the loss function of the generator and the discriminator. The loss function L_G of the generator is defined as Equation (20):

$$L_G = \sum L_{G_i} + \lambda_1 L_{ca} + \lambda_2 L_{DAMSM} \tag{20}$$

Among them, G_i represents the total loss function for the i -th generator. On the right-hand side of the equation, the first term is the generation loss function, the second term is the condition enhancement loss function, and the third term is the Deep Attentional Multi-modal Similarity Model (DAMSM) loss function [18]. Among them, λ_1 and λ_2 are the corresponding weights for the condition enhancement loss and the DAMSM loss, respectively, set to 1.0 and 0.5. The selection of these hyperparameters is based on sensitivity analysis, where the values of λ_1 and λ_2 are adjusted under different datasets and experimental settings to observe changes in model performance. Ultimately, these values are determined to effectively enhance the quality and semantic consistency of the generated images while ensuring the convergence speed of the model.

Generate the loss function L_{G_i} is expanded as Equation (22):

$$L_{G_i} = \frac{1}{2}\left[E_{x\sim p_{G_i}} \ln D_i(x) + E_{x\sim p_{G_i}} \ln D_i(x, s)\right] \tag{21}$$

Among them, D_i represents the i -th discriminator. Within the parentheses on the right-hand side of the equation, the first term is the unconditional loss function, whose function is to make the generated artistic design images more realistic, approaching the distribution of real images. The second term is the conditional loss function, whose function is to ensure semantic consistency between the generated artistic design images and the text descriptions [19].

The conditional enhancement loss function L_{ca} is expanded as Equation (22):

$$L_{ca} = D_{KL}\left(N(\mu(s), \sum(S) \,||\, N(0,1)\right) \tag{22}$$

Among them, $D_{KL}(\ )$ represents the formula for calculating the Kullback-Leibler (KL) divergence, which is used to measure the difference between two probability distributions. $\mu(s)$ and $\sum(S)$ are formulas for calculating the mean and variance, respectively. The function of L_{ca} is to enhance the diversity of the training data by introducing randomness, thereby preventing model overfitting and improving the model's generalization ability.

For the DAMSM loss function, its function is to measure the degree of match between artistic design images and text through pre-trained text encoders and image encoders, ensuring semantic consistency between the text and the corresponding artistic design images [20]. Specifically, the DAMSM loss function guides the generator to produce images that highly match the text descriptions by calculating the similarity between text features and image features.

The loss function $L_{D_{total}}$ of the discriminator is Equation (23):

$$L_{D_{total}} = \sum_i \left(L_{D_i} + L_{CD_i}\right) \tag{23}$$

Among them, L_{D_i} represents the unconditional loss function, which is used to discriminate whether the input image is real or fake, i.e., to determine whether the image is generated by the generator or comes from the real dataset. L_{CD_i} represents the conditional loss function, which is used to discriminate whether the input image and the text description are semantically consistent. These two loss functions jointly act on the discriminator, enabling it to make accurate judgments between real and generated images, as well as between images and their corresponding text descriptions. L_{D_i} and L_{CD_i} are expanded as Equations (24) and (25):

$$L_{D_i} = -\frac{1}{2}\left[E_{x\sim p_{data}} \ln(D_i(x)) + E_{x\sim G_i} \ln(1 - D_i(x))\right] \tag{24}$$

$$L_{G_i} = -\frac{1}{2}\left[E_{x\sim p_{G_i}} \ln(D_i(\mathrm{s}, x)) + E_{x\sim G_i} \ln(1 - D_i(\mathrm{s}, x))\right] \tag{25}$$

In the aforementioned formulas, the selection of weights λ_1 and λ_2 for the various loss function components significantly impacts the model's performance. Through sensitivity analysis, it has been found that $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$ can effectively enhance the quality and semantic consistency of the generated images while ensuring stable training of the model. The determination of these weights is based on experimental validation and the evaluation of the model's performance. With this, the process of generating artistic design images from text is completed.

3 Experimental analysis

In order to verify the application effect of the cross-modal attention based text generation art design image method proposed in this paper, the experimental environment is constructed as follows: Ubuntu 16.04, CPU is i7-4790k, GPU is GeForce GTX 1080Ti, and the experimental code uses the Pytorch deep learning framework, which is run on the GPU.

The dataset used is manually constructed, and the art and design images in the dataset are mainly from two online art platforms, Artstor (Artstor) and DeviantArt (DeviantArt). Among them, Artstor is a digital library that provides services to educational institutions and has a collection of more than 2 million images of art and cultural heritage from all over the world. Artstor is an invaluable resource for machine learning projects and research, especially in visual analytics, art style learning, and image recognition, etc. DeviantArt is an art community platform with a large database of artworks, including paintings, photography, and digital art. Although this platform may not be as open as some museums' official resources, it provides a diverse and modern selection of artworks suitable for researchers who are interested in modern art styles and trends. To this end, this paper selects 6000 art and design images from each of the two platforms and constructs a dataset containing 12000 images. In the data preprocessing stage, resolution normalization was performed on each image to ensure consistency in the input model data; At the same time, 8 text descriptions that match each image were segmented using a simple segmentation method based on spaces and punctuation, converting the text into a word vector form that the model can handle. Finally, 8400 images were used as the training set, and the remaining 3600 images were used as the testing set for subsequent experimental validation.

In this paper, we apply the cross-modal attention based generative adversarial network model (CMAGAN) parameter settings are shown in Table 2.

Table 2: Parameter Settings of CMAGAN Model

Parameter Name	Set Value
Text Encoder Type	Pre trained LSTM
Dimension of sentence feature vector	256
Dimension of word feature vector	18×300
Repetition frequency of content aware upsampling module	4

Upsampling rate	2
Adaptive Convolutional Kernel Prediction Module Convolutional Layer Size	3×3
Content encoder size	512×8×8
Channel attention convolution module global average pooling	Application
Size of channel attention weight matrix	64
Spatial attention module perception layer mapping dimension	256
Channel Attention 1×1 Convolutional Block	Application
Dynamic memory attention unit writing entry control parameters	1×256, 1×64 matrices
Dynamic Memory Attention Unit MoM Module	Apply self attention mechanism
Generator loss function weight	$\lambda 1 = 1.0$, $\lambda 2 = 0.5$
Discriminator loss function type	Combination of unconditional loss and conditional loss
Learning rate	0.0002
Batch Size	64
Training epochs	100

In Table 2, the text encoder type is selected as pre trained LSTM, which is used to encode text descriptions into sentence feature vectors (dimension 256) and word feature vectors (dimension 18×300). The content aware upsampling module has a repetition frequency of 4 times, with each upsampling rate of 2. The image resolution is gradually improved through an adaptive convolutional kernel prediction module (convolutional layer size of 3×3) and a content encoder (size of 512×8×8). The channel attention convolution module adopts global average pooling and sets the size of the channel attention weight matrix to 64 to enhance the channel information of the feature map. The spatial attention module sets the dimension to 256 through perceptual layer mapping and applies 1 × 1 convolutional block for channel attention adjustment. The MoM module of dynamic memory attention unit applies self attention mechanism, which specifically includes controlling the writing of entries through 1×256 and 1×64 matrices, dynamically weighting and fusing features using self attention mechanism, thereby enhancing the expression ability of visual features and making the generated images more semantically and visually rich and realistic. The weight of the generator loss function is set to $\lambda 1=1.0$ and $\lambda 2=0.5$ to balance the effects of different loss terms. The discriminator loss function type combines unconditional loss and conditional loss to more comprehensively evaluate the quality of generated images.

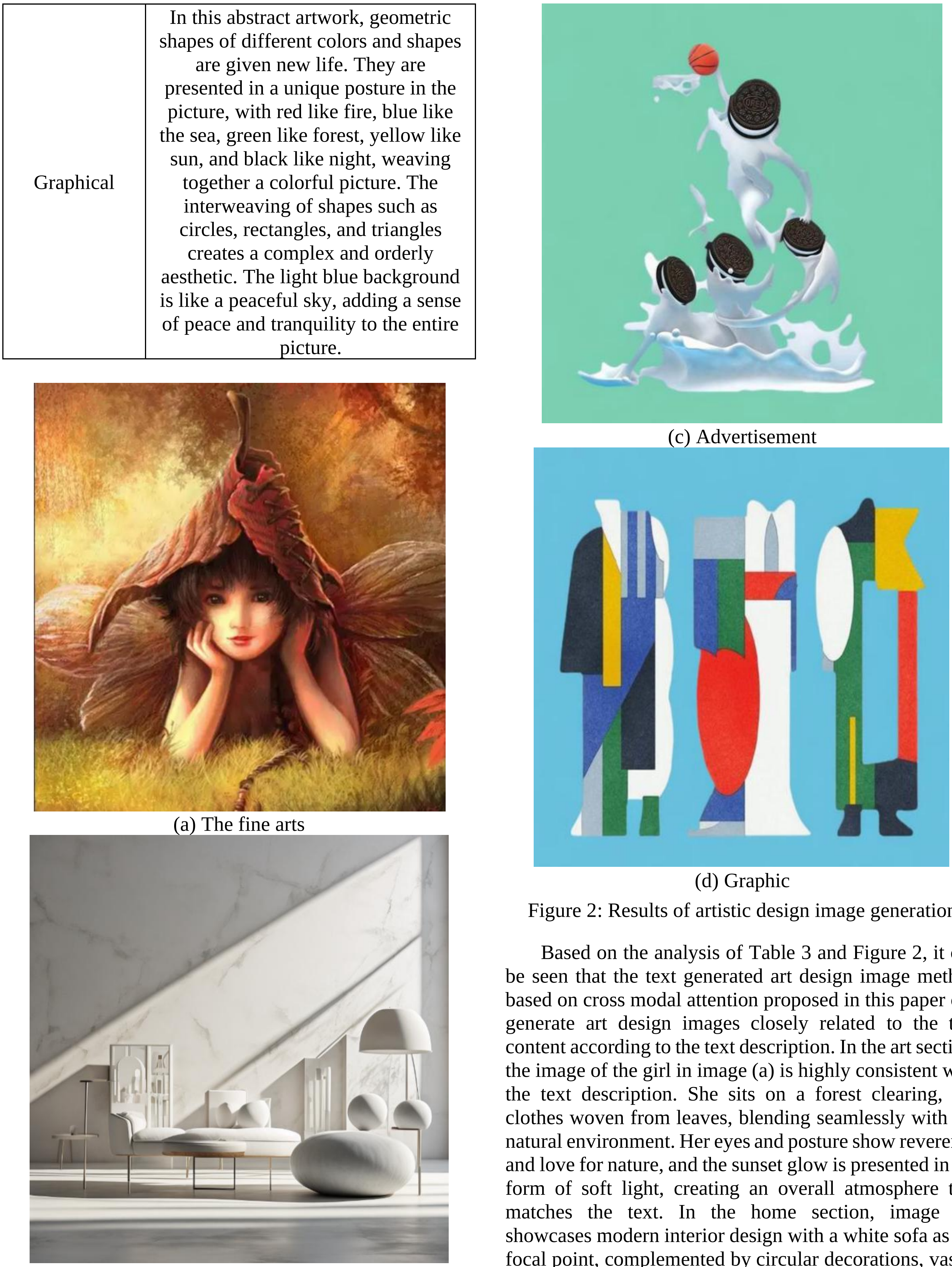
Based on the parameters shown in Table 2, the model in this paper is trained using the standard training process of generative adversarial networks. Improve model performance by alternately optimizing the generator and discriminator. During training, use a batch size of 64 samples for small batch gradient descent, set the learning rate to 0.0002, and continue training for 100 cycles. During the training process, normalization is performed in the data preprocessing stage to ensure stability. To prevent overfitting, data augmentation is used to expand the diversity of the training set. An early stopping strategy is implemented to terminate the training in advance based on the performance of the validation set, and weight decay (L2 regularization term) is introduced in the optimizer to ensure that the generated art design images are semantically consistent with the text description, thereby improving image clarity and diversity.

In order to verify the application effect of the method in this paper, four types of text descriptions, namely, art, home, advertisement and graphics, are randomly selected from the test set, as shown in Table 3.

Table 3: Text description details

Text Description Type	Text Description Details
Fine Arts	The afterglow of the sunset fell on a girl sitting in a clearing in the forest, her clothes woven from delicate leaves that perfectly blended with the surrounding natural environment. Her eyes revealed reverence and love for nature, and her hands gently lifted her cheeks, as if listening to the breathing of the forest.
Home Furnishing	In this modern interior design, a white sofa has become the soul of the space. The circular decoration next to it is like a brilliant pearl, embellishing the entire space. The vases and books on the small table not only beautify the environment, but also reveal the owner's cultural heritage. The floor lamp adds a sense of fashion to the room with its unique design. The sunlight shining through the window makes the entire room appear brighter and more spacious, creating a comfortable and pleasant atmosphere.
Advertisement	This illustration features Oreo cookies as the main character, creating a fun filled basketball world. The Oreo biscuit on the left, dressed in a white "battle robe" and holding a basketball, looks like a commander on the court. The two figures in the middle and on the right interpret the charm of basketball in a jumping posture, and the black design of their heads makes them look even more unique. The milk splash below seems to be a witness to their intense confrontation. The light green background adds a fresh touch of color to this Oreo basketball game.

Graphical	In this abstract artwork, geometric shapes of different colors and shapes are given new life. They are presented in a unique posture in the picture, with red like fire, blue like the sea, green like forest, yellow like sun, and black like night, weaving together a colorful picture. The interweaving of shapes such as circles, rectangles, and triangles creates a complex and orderly aesthetic. The light blue background is like a peaceful sky, adding a sense of peace and tranquility to the entire picture.



(a) The fine arts

(b) Home Furnishing

(c) Advertisement

(d) Graphic

Figure 2: Results of artistic design image generation

Based on the analysis of Table 3 and Figure 2, it can be seen that the text generated art design image method based on cross modal attention proposed in this paper can generate art design images closely related to the text content according to the text description. In the art section, the image of the girl in image (a) is highly consistent with the text description. She sits on a forest clearing, her clothes woven from leaves, blending seamlessly with the natural environment. Her eyes and posture show reverence and love for nature, and the sunset glow is presented in the form of soft light, creating an overall atmosphere that matches the text. In the home section, image (b) showcases modern interior design with a white sofa as the focal point, complemented by circular decorations, vases, books, and uniquely designed floor lamps, creating a bright, spacious, and comfortable atmosphere that is consistent with the text description. In the advertising section, image (c) depicts an Oreo biscuit as the protagonist, creating a basketball world. The biscuit character wears a white "battle robe" and holds a basketball. The splashing milk and light green tones in the background add fun and vitality, which is consistent with the basketball world theme and details described in the text. In terms of graphics, image (d) presents an abstract artwork, with geometric shapes of different colors and shapes intertwined to form a complex and orderly beauty. The light blue background echoes the peaceful sky described in the text, adding peace and tranquility to the picture. These visual elements vividly convey the emotions and themes in the text description through the use of details and colors, demonstrating the excellent effects of this method in presenting details, creating atmosphere, and overall aesthetics.

The quality of the generated images was evaluated with respect to diversity and semantic consistency using three evaluation criteria:

(1) R-precision. It is used to evaluate the semantic consistency between the generated image and the input text conditions. The higher the R value, the higher the semantic consistency between the generated image and the input text condition.

(2) Initial Score (Inception Score, IS). Used to measure the clarity and diversity of the generated images, it is the quality of the generated images by calculating the relative entropy loss of the marginal and conditional distributions, as shown in Equation (26):

$$I_{IS} = \exp(E_x D_{KL} p(y \mid x) \parallel p(y)) \quad (26)$$

Of which: x denotes the sample generated by the generator; $p(y)$ denotes the marginal distribution; $p(y \mid x)$ represents the distribution obtained by the x input image classification network; $D_{KL}(A \parallel B)$ represents the KL divergence between A and B, is used to measure the similarity between the two distributions, the larger the *IS* value, the higher the quality of the generated image.

(3) Frechet Inception Distance Score (FID). It is used to calculate the distance between the generated image and the real image distribution as shown in Equation (27):

$$F_{FID} = \left(\tau_x - \tau_{\hat{x}}\right)^2 + Tr(\upsilon_x + \upsilon_{\hat{x}} - 2\sqrt{\upsilon_x \upsilon_{\hat{x}}}) \quad (27)$$

Among them, τ_x and $\tau_{\hat{x}}$ respectively represent the feature mean of real art design images and generated art design images; υ_x and $\upsilon_{\hat{x}}$ respectively represent the covariance between real art design images and generated art design images.

Taking the above three indicators as the measurement standard of the application model of this method, the ablation experiment is carried out to generate four artistic design images, namely, art, furniture, advertising and graphics. In the ablation experiment, model 1 only adopts the initial generation stage of images across modes, model 2 adds the initial thinning stage of images on the basis of model 1, and model 3 adds the re-thinning stage of images on the basis of model 2. The results of ablation experiments are shown in Table 4.

Table 4: Evaluation results of ablation experiments

Experimental Model	Art and Design Image Types	R-value Accuracy	Initial Score (IS)	Frechet Inception Distance Score (FID)
Model 1	Fine Arts	0.68	2.95	36.4
	Furniture	0.72	3.11	32.8
	Advertisement	0.65	2.81	39.2
	Graphical	0.72	3.05	34.6
Model 2	Fine Arts	0.75	3.32	30.2
	Furniture	0.78	3.45	28.6
	Advertisement	0.76	3.19	35.4
	Graphical	0.76	3.35	31.2
Model 3	Fine Arts	0.86	3.65	20.4
	Furniture	0.85	3.88	24.8
	Advertisement	0.78	3.47	22.8
	Graphical	0.88	3.69	19.4

According to the analysis of the data in Table 4, as the model architecture gradually optimizes from Model 1 to Model 2, and then to Model 3, the generation effect of art and design images shows a significant improvement trend. Specifically, Model 2 adds a preliminary image refinement stage based on Model 1. By introducing spatial attention and channel attention mechanisms, the model can more accurately match text and image features, thereby improving the quality of generated images; Model 3 further adds an image refinement stage based on Model 2, utilizing the secondary attention (AoA) mechanism to pay deeper attention to and fuse image features, significantly enhancing the expressive power of visual features. This optimization process is not only reflected in the gradual improvement of R-value accuracy (for example, the R-value accuracy of graphic design images has increased from 0.72 to 0.88, indicating that the model can more accurately capture the semantic information described in text), but also in the increase of initial score (IS) (for example, the IS value of art images has increased from 2.95 to 3.65, reflecting the improvement of generated image diversity and quality) and the decrease of Frechet Inception distance score (FID) (for example, the FID score of advertising images has decreased from 39.2 to 22.8, indicating a reduction in the distribution distance between generated images and real images, and an enhancement of realism). In summary, the ablation experiment confirmed the crucial role of the initial image refinement stage and the image re refinement stage (especially the AoA mechanism) in model performance. Model 3 not only generates images that are highly consistent with the input text, but also ensures high quality and diversity of the images, while being closer to the distribution distance of the real images.

Based on 'LOVE', use AI to generate art design images as test raw images, as shown in Figure 3.

Figure 3: Art and design image generation

In order to verify the advantages of this method in the art design of text generation compared with the existing research, the methods studied by Endo Y, Bahani M et al., Nezhad N M et al. and Watanabe Y et al. are selected as the comparison, and the visual semantic similarity (VSS) is selected as the evaluation index of the image generation effect, in which VSS evaluates the consistency between the generated image and the input text by comparing their distance or similarity in the semantic space. This usually involves mapping images and texts to a common semantic space, and then calculating the similarity score between them. The higher the score, the more semantically the generated image matches the input text. It should be noted that all presented VSS indicator results have undergone rigorous statistical significance testing (paired t-test) to ensure the reliability and effectiveness of performance differences between methods, thus more scientifically verifying the advantages of our method compared to other comparative methods. Taking Figure 3 "LOVE" as the original image, further compare the performance of the five methods, and the results are shown in Figure 4. The VSS index results of text generation art design images using five methods are shown in Figure 5.

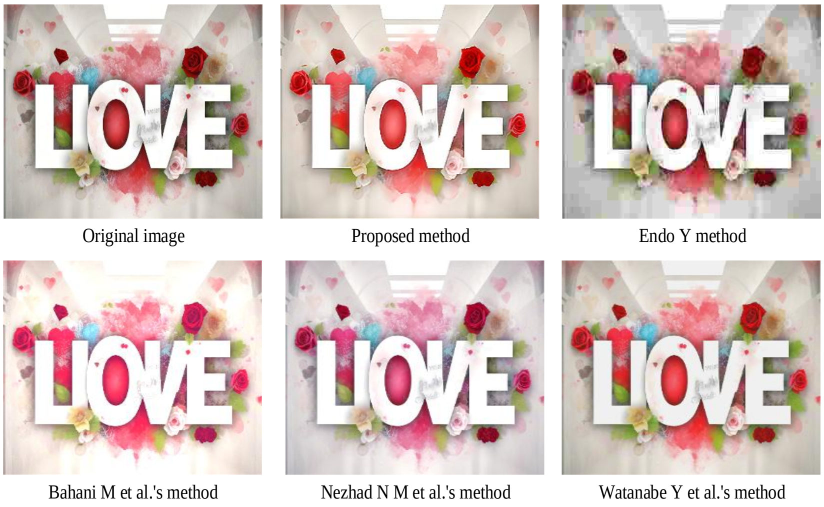

Original image
Proposed method
Endo Y method
Bahani M et al.'s method
Nezhad N M et al.'s method
Watanabe Y et al.'s method

Figure 4: Text-generated image visual comparison

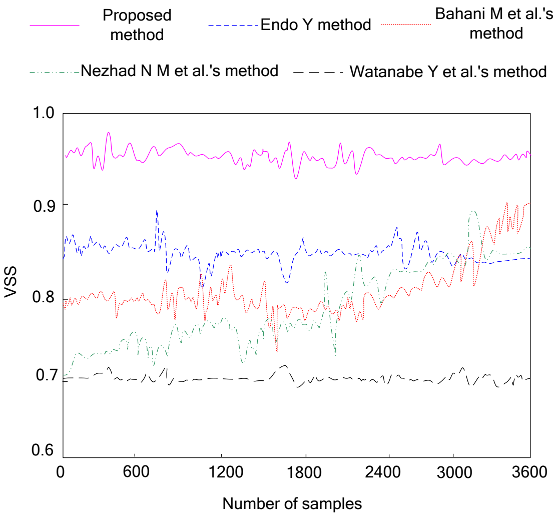


Figure 5: VSS index results of text generation art design images using five methods

From Figure 4, it can be seen that the original image has bright colors and rich details, the word "LOVE" is clear and three-dimensional, and decorative elements such as background flowers are lifelike. The method proposed in this article generates images that are closer to the original image in terms of color and detail, significantly improving image quality and preserving most visual features, especially in terms of detail preservation and clarity. In contrast, the images of Endo's [3] method are slightly blurry and have some loss of details; Although Bahani et al.'s [4] method has a relatively balanced overall effect, the sense of three dimensionality is slightly inferior and there is blurring in some areas; The overall visual effect of Nezhad et al.'s [5] method is close to the original image, but some details are significantly lost; Watanabe et al.'s [6] method performs well in color softness, but lacks clarity in some details. Overall, the method proposed in this paper outperforms other methods in terms of visual effects, better balancing detail preservation and image quality improvement, thus verifying the qualitative evaluation of its effectiveness in the manuscript. The experimental results in Figure 5 show that the method proposed in this paper exhibits better VSS scores than other comparative methods under multiple sample sizes. Specifically, as the number of samples increases, although the VSS scores of all methods improve, the improvement of this paper's method is the most significant and always remains at a high level. In contrast, the methods of Endo [3], Bahani et al. [4], Nezhad et al. [5] and Watanabe et al. [6] show some performance under certain sample sizes, but their VSS scores are lower than that of this paper's method in general. This result fully verifies the advantage of this paper's method in text generation of art design image effects. By adopting more advanced algorithms and model structures, this method is able to understand the semantic information of the input text more accurately and transform it into images with artistic design effectively. This not only improves the semantic consistency between image and text, but also provides users with richer and more diverse visual experiences.

In order to evaluate the performance of the proposed method in terms of computational complexity, comparative experiments were conducted with the baseline model (traditional DCGAN) and the methods proposed by Endo [3], Bahani et al. [4], Nezhad et al. [5], Watanabe et al. [6]. The experiment mainly focuses on the inference time, training time, and resource requirements (such as GPU hours) of the model. Among them, the computational complexity selected three indicators: inference time (the average time required for the model to generate an art design image), training time (the total time required for the model to train and converge), and GPU hours (the total number of GPU usage hours during the model training process, reflecting resource requirements). All experiments were conducted in the same hardware environment (Ubuntu 16.04, CPU i7-4790k, GPU GeForce GTX 1080Ti) and software environment (Pytorch deep learning framework) to ensure fairness and comparability of results. The experimental results are shown in Table 5.

Table 5: Calculation complexity analysis

Method	Inference time (seconds/sheet)	Training time (s)	GPU hours
Baseline model	0.8	120	400
Endo [3]	1.2	150	500
Bahani et al. [4]	1.0	130	450
Nezhad et al. [5]	1.5	180	600
Watanabe et al. [6]	1.1	140	480
Proposed method	1.3	160	550

According to Table 5, in terms of inference time, using the traditional DCGAN as the baseline model, its inference time is the shortest, at 0.8 seconds per image, while the inference time of our method is 1.3 seconds per image. Although slightly longer than the baseline model, it is shorter than other comparative methods proposed by Endo [3], Bahani et al. [4], Nezhad et al. [5], and Watanabe et al. [6]. This indicates that although CMAGAN introduces more attention mechanisms in generating art design images, it does not significantly increase inference time. In terms of training time, the traditional DCGAN baseline model has the shortest training time, which is 120 seconds, while the CMAGAN model has a training time of 160 seconds, which is longer than the baseline model, but at a moderate level compared to other comparative methods. This may be due to the relatively complex structure of the CMAGAN model, which requires more training time to converge. In terms of GPU hours, the traditional DCGAN baseline model has the lowest GPU hours at 400 hours, while the CMAGAN has 550 hours, which is more than the baseline model. However, compared with other comparison methods, it is still competitive, indicating that although CMAGAN requires more GPU resources during the training process, it has not reached the level of other methods. Based on the above analysis, the method proposed in this paper performs well in terms of computational complexity. Although the inference time and training time are slightly longer than the baseline model, it has advantages compared to other comparative methods. Moreover, CMAGAN can maintain high quality and diversity when generating art and design images, and is semantically consistent with text descriptions. This is due to its unique cross modal attention mechanism and phased refinement strategy. Therefore, CMAGAN achieves a good balance between computational complexity and generation performance.

To further evaluate the performance of our method in text to art design image generation tasks, we compared the five methods in Table 5 in an environment with more diverse and larger datasets. During the experiment, a more diverse and larger dataset containing over 50000 pairs of text and art design images was constructed. The text descriptions covered a wide range of topics from abstract concepts to concrete scenes, while the image styles spanned multiple fields such as modern art, retro design, and future technology to ensure that the model could learn rich visual and semantic associations, thereby improving the quality and diversity of generated images. At the same time, human evaluation has been introduced to more intuitively reflect the realism of generated images and the alignment of text images. The results are shown in Table 6.

Table 6: Realism of images and alignment of text images

Method	Human Assessment - Realistic Rating	Human Assessment - Text Image Alignment Score

Baseline model	3.5/5	3.2/5
Endo [3]	3.8/5	3.5/5
Bahani et al. [4]	3.6/5	3.3/5
Nezhad et al. [5]	3.7/5	3.4/5
Watanabe et al. [6]	3.4/5	3.1/5
Proposed method	4.2/5	4.0/5

According to Table 6, the images generated by our method perform excellently, achieving the highest realism score (4.2/5) and text image alignment score (4.0/5), indicating that users consider these images to be more realistic and highly match the text description. To gain a deeper understanding of user experience, we invited 50 professionals and enthusiasts in the field of art and design to rate and comment on the images generated by each method. Most users reported that the images generated by CMAGAN were the best in terms of realism and text image alignment, with rich details, bright colors, and accurate reflection of text content. However, the images generated by the baseline model and other comparison methods had problems with blurring, distortion, or inconsistency with text descriptions. At the same time, users generally appreciated the ability of this method in generating diverse art and design images, believing that its images have higher creativity and artistic value. The effectiveness of CMAGAN in text to art design image generation tasks has been verified through comparative experiments and user research. This confirms that the method proposed in this article can generate art and design images that are more realistic and highly matched with textual descriptions. These results not only demonstrate the technical strength of CMAGAN, but also provide strong support for its application in the field of art and design.

In order to explore the impact of different text encoders on the task of generating art design images from text and verify the robustness of our method (CMAGAN), two different types of text encoders were used: the Transformer based CLIP model and the traditional LSTM model. The experimental input includes different types of adversarial text descriptions, including clear, fuzzy, and contradictory text inputs, to test the performance of the model under complex text conditions. The experiment compared the performance of our method using different text encoders, with evaluation metrics including R-value accuracy, initial score (IS), and Frechet Inception distance score (FID), to comprehensively analyze the quality, diversity, and semantic consistency of the generated images with the text description. The results are shown in Table 7.

Table 7: Impact of different text encoders on the task of generating artistic design images from text

Text Encoder	Text description type	R-value accuracy	Initial Score (IS)	Frechet Inception Distance Score (FID)
CLIP	Clear	0.86	3.65	20.4
	Vague	0.78	3.20	28.5
	Contradiction	0.72	2.95	35.2
LSTM	Clear	0.82	3.40	25.6
	Vague	0.70	2.80	32.8
	Contradiction	0.65	2.50	40.5
Method described in this paper (CMAGAN+CLIP)	Clear	0.88	3.70	19.8
	Vague	0.80	3.30	26.0
	Contradiction	0.75	3.00	30.1
The method described in this paper (CMAGAN+LSTM)	Clear	0.84	3.50	23.2
	Vague	0.74	3.00	29.5
	Contradiction	0.68	2.60	37.8

According to Table 7, both CLIP and LSTM as text encoders can provide effective text features for CMAGAN under clear text descriptions. However, CLIP performs slightly better in R-value accuracy, IS, and FID, demonstrating its advantages in capturing semantic information in text; Under fuzzy or contradictory text descriptions, the robustness of CLIP is significantly stronger than LSTM. The CMAGAN using CLIP as a text encoder has a smaller decrease in various evaluation indicators, indicating that it can better handle complex or uncertain text information. No matter which text encoder is used, CMAGAN can generate high-quality art and design images that are highly consistent with the text description under clear text description, and can still maintain certain generation performance under fuzzy or contradictory text description. Especially when using CLIP, the robustness of CMAGAN is significantly improved. CLIP, as a Transformer based model, performs outstandingly in capturing semantic information of text and processing complex texts. Compared with LSTM, it can provide richer and more accurate text features, thereby assisting CMAGAN in generating higher quality and more consistent art and design images with text descriptions.

4 Discussion

The cross-modal attention based generative adversarial network model proposed in this study has achieved significant results in the task of text generation of art and design images, demonstrating clear advantages compared to previous work. Through ablation experiments and comparisons with the research methods of Endo [3], Bahani et al. [4], Nezhad et al. [5], Watanabe et al. [6], the CMAGAN model performed well in evaluation metrics such as R-value accuracy, initial score (IS), Frechet Inception distance score (FID), and visual semantic similarity (VSS).

The reason why the CMAGAN model outperforms other models is mainly attributed to its unique cross modal attention mechanism and phased refinement strategy. In the initial generation stage of cross modal images, the model encodes text descriptions through pre trained RNNs and generates initial art design images using content aware upsampling modules and channel attention convolution modules, effectively capturing the correlation between text and images. Subsequently, in the initial and further refinement stages of the image, the model introduces spatial attention, channel attention mechanisms, and quadratic attention (AoA) mechanisms to further pay attention to and fuse image features, enhancing the expression of visual features. This phased refinement strategy enables the model to gradually optimize the quality of generated images and improve semantic consistency between images and text.

In addition, the CMAGAN model also takes into account the quality, diversity, and similarity with real images when generating images through the comprehensive loss function, thereby generating artistic design images that not only conform to textual descriptions but also have high quality and rich details.

However, despite the significant achievements of the CMAGAN model in text generated art design image tasks, there are still some potential failure cases. For example, overly complex or vague text descriptions may hinder the model’s ability to accurately interpret semantic information. This can result in generated images deviating from the original text description. In addition, since the model is trained on a large amount of data, if there are biases or deficiencies in the training data, it may also affect the generation performance of the model. Therefore, in future research, it is necessary to further explore how to optimize the text comprehension ability of the model and how to improve the generation performance of the model under complex and fuzzy text descriptions. At the same time, it is also necessary to pay attention to the quality and diversity of training data to ensure that the model can generate more accurate and rich artistic design images.

5 Conclusion

In this paper, we propose a cross-modal attention-based method for generating art design images from text, and verify the practical application of this method through a series of experiments. The experimental results show that the method is able to generate art design images that match the content of the text based on the text description, and it shows better results in detail presentation, atmosphere creation and overall aesthetics. And through the R-value accuracy, initial score (IS) and Frechet Inception Distance Score (FID) three kinds of evaluation criteria by way of ablation experiments to verify that the model applied in this paper's method can obtain the advantage of more significant image generation results. The advantages of the proposed method in generating artistic design images from text are conclusively validated through comparative analysis with existing literature. This paper's method can more accurately understand the semantic information of the input text and effectively transform it into images with artistic design sense, which provides users with richer and more diverse visual experiences.

References

[1] Tuerhong, G., Tian, L., Wushouer, M., & Dai, X. (2024). An end-to-end image-text matching approach considering semantic uncertainty. Neurocomputing, 607, 128386. DOI: 10.1016/j.neucom.2024.128386.

[2] Li, W., Kong, D. Q., Cao, G. G. Dai, C. Y. (2022). Cross-modality Medical Image Registration Based on Attention Residual Network. Computer Simulation, 9(11), 224-229. DOI: 10.3969/j.issn.1006-9348.2022.11.044.

[3] Endo, Y. (2024). Masked-attention diffusion guidance for spatially controlling text-to-image generation. The visual computer, 40(9), 6033-6045. DOI:10.1007/s00371-023-03151-y

[4] Bahani, M., El Ouaazizi, A., & Maalmi, K. (2023). The effectiveness of T5, GPT-2, and BERT on text-to-image generation task. Pattern recognition letters, 173, 57-63. DOI: 10.1016/j.patrec.2023.08.001

[5] Nezhad, N. M., Mirtaheri, S. L., & Shahbazian, R. (2023). Popular image generation based on popularity measures by generative adversarial networks. Multimedia Tools & Applications, 82(14), 20873-20897. DOI:10.1007/s11042-022-14090-6

[6] Watanabe, Y., Ogawa, N., Maeda, K., Ogawa, T., & Haseyama, M. (2024). Automatic findings generation for distress images using in-context few-shot learning of visual language model based on image similarity and text diversity. Journal of Robotics & Mechatronics, 36(2), 353-364. DOI:10.20965/jrm. 2024.p0353

[7] Thakur, N., Khan, N. U., & Sharma, S. D. (2023). A two-phase ultrasound image de-speckling framework by nonlocal means on anisotropic diffused image data. Informatica: An International Journal of Computing and Informatics, 47(2), 221-234. DOI:10.31449/inf. v47i2.4378

[8] Mehmood, R., Bashir, R., & Giri, K. J. (2024). VTM-GAN: video-text matcher based generative adversarial network for generating videos from textual description. International Journal of Information Technology, 16(1), 221-236. DOI:10.1007/s41870-023-01468-4

[9] Helmy, T., & Djatmiko, F. (2023). Framework for automatic semantic annotation of images based on image's low-level features and surrounding text. Arabian journal for science and engineering,48(2), 1991-2007. DOI:10.1007/s13369-022-06828-z

[10] Phaphuangwittayakul, A., Ying, F., & Chakpitak, Z. N. (2023). Few-shot image generation based on contrastive meta-learning generative adversarial network. The visual computer, 39(9), 4015-4028. DOI:10.1007/s00371-022-02566-3

[11] Kabir, A. I., Mahomud, L., Ald, A., & Ahmed, R. (2024). Empowering local image generation: harnessing stable diffusion for machine learning and AI. Informatica Economica, 28(1), 25-38. DOI:10.24818/issn14531305/28.1.2024.03

[12] Mishra, A., Agrawal, A., & Bhasker, S. (2024). Hybrid explainable image caption generation using image processing and natural language processing. International Journal of System Assurance Engineering and Management, 15(10), 4874-4884. DOI:10.1007/s13198-024-02495-5

[13] Nemer, Z. N., Jasim, W. N., & Harfash, E. J. (2023). Implementation of multiple cnn architectures to classify the sea coral images. Informatica: An International Journal of Computing and Informatics, 47(1), 43-50. DOI:10.31449/inf. v47i1.4429

[14] Praveen, R. G., Cardinal, P., & Granger, E. (2023). Audio–visual fusion for emotion recognition in the valence–arousal space using joint cross-attention. IEEE Transactions on Biometrics, Behavior, and Identity Science, 5(3), 360-373. DOI:10.1109/TBIOM.2022.3233083

[15] Pourkeshavarz, M., Nabavi, S., Moghaddam, M. E., & Shamsfard, M. (2024). Stacked cross-modal feature consolidation attention networks for image captioning. Multimedia Tools & Applications, 83(4), 12209-12233. DOI:10.1007/s11042-023-15869-x

[16] Tomer, M., & Kumar, M. (2022). STV-BEATS: skip thought vector and bi-encoder based automatic text summarizer. Knowledge-Based Systems, 240, 108108. DOI: 10.1016/j.knosys.2021.108108

[17] Mocanu, B., Tapu, R., & Zaharia, T. (2023). Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning. Image and vision computing, 133, 104676. DOI: 10.1016/j.imavis.2023.104676

[18] Matú Dopiriak, Gerec, J., & Gazda, J. (2024). Reconstruction of photorealistic 3d urban scenes using radiance fields as digital twins for autonomous driving. Acta Electrotechnica et Informatica, 24(4), 27-34. DOI:10.2478/aei-2024-0015

[19] Xu, M., Rahman, H. A., & Li, F. (2023). Automated generation of Chinese text-image summaries using deep learning techniques. Traitement du Signal, 40(6), 2835-2843. DOI:10.18280/ts.400644

[20] Borra, S. R, Ritika, K., Reddy, N. A., Shanvitha, N., Anusha, T., & Rajeswari, O. S. S. R. (2024). Parkinson net: convolutional neural network model for parkinson disease detection from image and voice data. Informatica (03505596), 48(2). DOI:10.31449/inf. v48i2.5077