applied sciences

MDPI

Article

Image Generation with Global Photographic Aesthetic Based on Disentangled Generative Adversarial Network

Hua Zhang [1,2], Muwei Wang [1], Lingjun Zhang [1,3,*], Yifan Wu [1] and Yizhang Luo [1]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China; zhangh@hdu.edu.cn (H.Z.); wangmuwei1998@163.com (M.W.); yfwu@hdu.edu.cn (Y.W.); luoyizhang@hdu.edu.cn (Y.L.)

[2] Key Laboratory of Network Multimedia Technology of Zhejiang Province, Zhejiang University, Hangzhou 310018, China

[3] Key Laboratory of Brain Machine Collaborative Intellignece of Zhejiang Province, Hangzhou Dianzi University, Hangzhou 310018, China

* Correspondence: zhanglingjun@hdu.edu.cn

Abstract: Global photographic aesthetic image generation aims to ensure that images generated by generative adversarial networks (GANs) contain semantic information and have global aesthetic feelings. Existing image aesthetic generation algorithms are still in the exploratory stage, and images screened or generated by a computer have not yet achieved relatively ideal aesthetic quality. In this study, we use an existing generative model, StyleGAN, to build the height of image content and put forward a new method based on the GAN disentangled representation of a global aesthetic image generation algorithm by mining GANs' latent space, potential global aesthetic feeling, and aesthetic editing of the original image to realize the aesthetic feeling and content of high-quality global aesthetic image generation. In contrast with the traditional aesthetic image generation methods, our method does not need to retrain GANs. Using the existing StyleGAN generation model, by learning a prediction model to score the generated image and the score as a label to learn a support vector machine decision surface, we use the learned decision to edit the original image to obtain an image with a global aesthetic feeling. This method solves the problems of poor content construction effect and poor global beauty of the aesthetic images generated by the existing methods. Experimental results show that the proposed method greatly increases the aesthetic score of the generated images and makes the generated images more in line with people's aesthetic.

Keywords: GANs; global photographic aesthetic; disentangled representation; image generation; generative adversarial network

check for updates

Citation: Zhang, H.; Wang, M.; Zhang, L.; Wu, Y.; Luo, Y. Image Generation with Global Photographic Aesthetic Based on Disentangled Generative Adversarial Network. *Appl. Sci.* **2023**, *13*, 12871. https://doi.org/10.3390/app132312871

Academic Editor: Yudong Zhang

Received: 9 October 2023
Revised: 7 November 2023
Accepted: 17 November 2023
Published: 30 November 2023

1. Introduction

Aesthetic image generation is widely used and has penetrated every aspect of human life. For example, in print advertising design with pictures, the image is required to be clear, real, and have a high degree of beauty to attract attention. Generative adversarial networks (GANs) are widely used in image generation and have developed rapidly in recent years; several high-quality derived models based on GANs have been proposed. These generative models have made great breakthroughs in the fields of image generation such as indoor and outdoor scenes, animals, and flowers. However, the images generated by adversarial networks often focus more on the reconstruction of the image content, without any consideration of aesthetic factors. The resulting images lack the aesthetic feeling recognized by the public, which leads to their application scenarios being limited by their aesthetic quality. Therefore, this study aims to generate globally aesthetic images.

From the perspective of image aesthetics, a photographic image can be divided into two parts: specific content and aesthetic presentation [1]. The generation effect of existing generative models on image content has reached the level of mixing the spurious with

genuine and its aesthetic presentation is not satisfactory. Figure 1 shows a bridge image generated by the StyleGAN generation model and one taken by human beings. The image generated by GANs and its semantic expression are very accurate. It is almost impossible to distinguish the true from the false image with the human eye. However, compared with the bridge images taken by humans, there is a significant gap in its aesthetic quality. When human beings select images, they often screen from massive images according to the aesthetic sense of the images. Although the existing image generation models have been able to produce highly realistic images in fixed scenes and in large quantities, there are still no extensive application scenarios due to the uneven aesthetic quality of the generated images. Therefore, the image generation of GANs should not only satisfy high-quality content reconstruction, but also pursue a better visual perception.

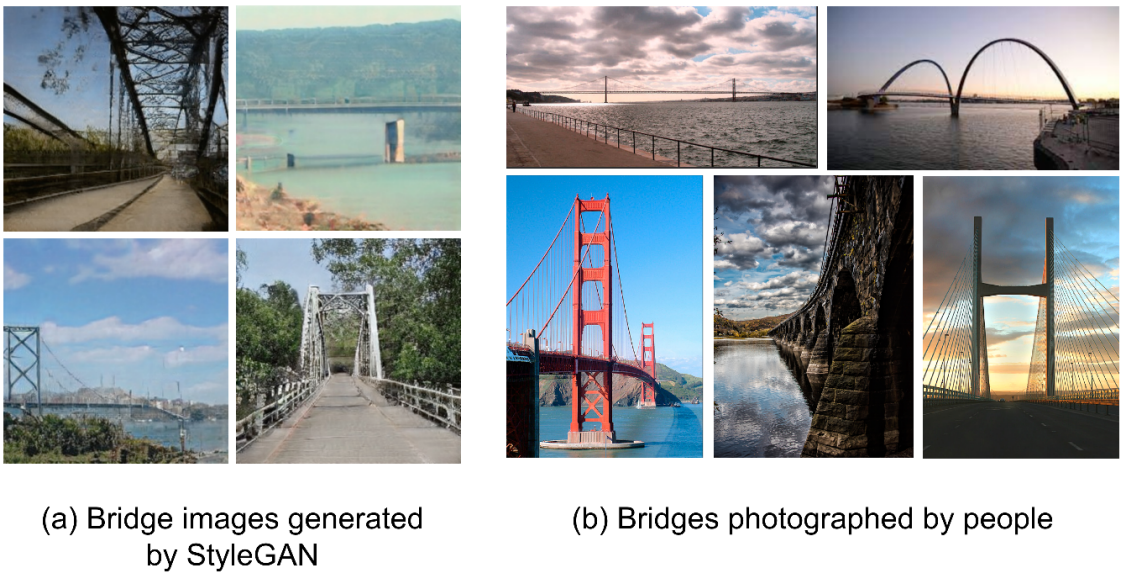

Figure 1. A comparison of the generated images and the images taken artificially.

GANs perform excellently in unconditional generation tasks, image processing, face editing, and other conditional generation tasks. However, their application in aesthetic image generation is still in the exploration stage and the existing global aesthetic image generation algorithm has the following problems:

- GAN models need to be redesigned and trained. Training GANs requires constructing appropriate loss function, evaluation indicators, and constraints to ensure the stability and effectiveness of the training, which is extremely difficult and resource-consuming. Compared with color, object, artistic style, and other attributes, the aesthetics of the image have strong uncertainty and subjectivity, and there is no basis for completely qualitative aesthetics; therefore; it is difficult to design and add appropriate aesthetic constraints for GANs to generate images with a better aesthetic effect.
- The current datasets used to train the aesthetic conditions of GANs are small, and the quality of the generated image content semantics is poor. The aesthetic condition training set of the GANs must meet the requirements of both semantic and aesthetic labels, and no large-scale dataset currently meets this requirement. Even with pre-training on larger datasets, the image semantics are less effective compared to those of the existing GANs trained on large-scale datasets.

To give a better aesthetic presentation effect to images generated by GANs, we proposed a method called global aesthetic image generation algorithm based on GAN's disentangled representation. It makes use of its ability to reconstruct high-quality image content and mine the global aesthetic disentangled representation in latent space to generate images with high overall aesthetic quality and realistic content. Since the algorithm is based on

the existing aesthetic image generation algorithm of GANs, there is no need to design the corresponding aesthetic loss and aesthetic constraints for the generation of an aesthetic image, and it is not necessary to redesign and train the generation model, which greatly saves on training time and training resources. Existing aesthetic image generation models require large-scale datasets with semantic labels and aesthetic labels.

The disentangled representation learning-based aesthetic GAN that we proposed differs from existing aesthetic image generation algorithms in many ways. First, our algorithm directly utilizes an existing GAN model for semantic image reconstruction, and on this basis, obtains both semantic and aesthetic images through aesthetic editing of the latent space. However, with existing aesthetic image generation algorithms, there is a need to design aesthetic losses and aesthetic constraints to train the conditional GAN models, which is very time-consuming and labor-intensive. Second, our algorithm does not need to be trained on a large-scale dataset with both semantic and aesthetic labels like an aesthetically conditioned GAN; we only need to train a prediction model on a dataset with only aesthetic labels, and then a support vector machine decision surface is learned based on the initial generated image and its aesthetic score. The desired image, with global aesthetics, can be obtained through the aesthetic editing of the latent space of the image.

In summary, this study makes the following contributions:

- We creatively apply the decoupled representation learning of GAN to aesthetic image generation. By learning the global aesthetic disentangled representation of the latent space, mining the aesthetic representation in the latent space of the StyleGAN model, and using the aesthetic representation learned to edit the latent space properly, we generate images with higher aesthetic quality.
- An aesthetic image generation algorithm based on an existing GAN model is designed to multiplex the high-quality reconstruction ability of the existing GAN model on the image content and add global aesthetic information to its latent space to make the generated image both aesthetic and semantic.
- By training the global aesthetic feeling prediction model to learn the global aesthetic feeling disentangled representation of GAN, the generation of images with more global aesthetic feeling is realized.
- The changes in aesthetic style in improving the global aesthetic feeling of different generative models are tracked and explored, and an experiment proves that the global aesthetic feeling of different scenes is closely related to a specific aesthetic style. It is of great significance to establish the mapping relationship between the image's global aesthetic feeling and the aesthetic style in the future.

This paper is divided into five sections. The Section 1 is the introduction, which mainly expounds the research background and research significance of this paper. First, the research content of global aesthetic feeling generation is described, and the main problems existing in this field are summarized. Second, we briefly explain the proposed algorithm and its difference from the existing methods for the problems encountered in the current research. In addition, the section also summarizes the contributions of this study and the organizational structure of this paper. The Section 2 summarizes the main work related to this paper. The Section 3 describes our proposed global aesthetic image generation algorithm based on GAN disentangled representations, including details of each step of the algorithm as well as training strategies and parameter settings. The Section 4 presents the experimental settings, as well as the basic analysis of different parameters of our method. We report on tests performed to compare our method with other methods. The section includes ablation studies and the visualization of some important "intermediate results". The Section 5 concludes the paper.

2. Related Work

In this chapter, we primarily focus on the analysis of aesthetic prediction models, the study of latent space of GANs, aesthetic image generation, and the datasets from the existing works.

2.1. Aesthetic Prediction Model

The evaluation of the aesthetic quality of images occupies an important position in computer vision. At present, some achievements have been made in the technology of image quality prediction [2]. A framework for using convolutional neural networks (CNNs) to predict the continuous aesthetic score of images is proposed, which can effectively evaluate the degree of aesthetic quality similar to the human system [3]. An image aesthetic prediction method based on weighted CNNs is proposed using a histogram prediction model to predict aesthetic scores, and to estimate the difficulty of aesthetic evaluation of the input images. A probabilistic quality representation method for deep blind image quality prediction proposed by Zeng et al. [4] retrains AlexNet and ResNetcnn to predict photo quality. Recently, a new multi-model recurrent attention convolutional neural network [5] was proposed, which consists of two streams: visual flow and language flow. The former uses a recurrent attention network to eliminate irrelevant information and focus on extracting visual features in some key areas. The latter uses Text-CNN to capture the high-level semantics of user comments. Finally, the multimodel decomposition bilinear pooling method is used to effectively integrate text features and visual features. In this study, we use Neural Image Assessment [6] as the image global aesthetic feeling prediction model. The model uses CNNs to predict the distribution of human opinion scores and is much simpler than other methods with similar performance. It is reliable in scoring images and highly correlated with human perception.

2.2. Study on Latent Space of GANs

A latent space is a space of a compressed representation of the data. The input variable z of GANs is unstructured, so it is proposed to decompose the latent variable into a conditional variable c and the standard input latent variable z. The decomposition of the latent space specifically includes both supervised methods and unsupervised methods. Typical supervised methods are CGAN [7] and ACGAN [8]. Recently, a latent space using supervised learning for GANs, which discovers more latent space about GAN by encoding the human future knowledge, was proposed [9]. Unsupervised methods do not use any label information and require disentanglement of the latent space to obtain meaningful feature representations. The disentangled representation of GANs is the process of separating the feature representation of the individual generating factors from the latent space. The disentangled representation can separate the explanatory factors of nonlinear interactions in real data, such as object shape, material properties, and light sources. The separation of properties can help researchers to manipulate GAN generation more intuitively. For the study of decoupled representation learning, Lee et al. [10] proposed an information distillation generation adversarial network, which learns separated representations based on vaa models and extracts the learned representations and additional interference variables into separate GAN-based generators for high-fidelity synthesis. In InterfaceGAN [11], the framework explains disentangled face representations learned by state-of-the-art GANs and deeply analyzes the properties of face semantics in latent space, detail the correlation between different semantics, and better disentangle them through subspace projections to provide more precise control over attribute manipulation. In this study, we solve the disentanglement problem of Z latent space using StyleGAN proposed by Karras et al. [12] StyleGAN combines eight fully connected layers to form a mapping network through which the Z of the original input is mapped to the W space. W is the same as the Z dimension but is more decoupled than the distribution of Z.

2.3. Aesthetic Image Generation

Aesthetic image generation refers to the generation of an image with aesthetic factors, so that the image has higher quality and is more in line with human aesthetics. Some traditional methods to improve the aesthetic quality of images, such as the super resolution reconstruction proposed by Li et al. [13], are to use the original image information to restore the super resolution image with clearer details and stronger authenticity. There is also an

image repair algorithm proposed by them [14], which performs well in the task of repairing irregular mask images, and the repair results have good performance in the aspects of edge consistency, semantic correctness, and overall image structure. These methods are closely related to improving image quality. To improve the quality of an image generated by GANs, some researchers studied the aesthetic image generation algorithm based on the conditional generation adversarial network, and attempted to design aesthetic losses and aesthetic constraints to train the aesthetic condition GANs, so that the generated images are both semantic and aesthetic [15]. Murray et al. proposed PFAGAN [16], a conditional GAN with aesthetics and semantics as dual labels, which combines the conditional aesthetic information and conditional semantic information as the training constraints, enabling the generative model to learn the content semantics and image beauty. Zhang et al. proposed a modified aesthetic condition GAN based on unsupervised representation learning with deep convolutional generative adversarial networks [17], where the network was trained with a batch size set to 256 for a total of 10,000 rounds of training iterations. There are still some problems in the existing methods of aesthetic image generation. The training of PFAGAN, which was proposed by Murray et al., is extremely time-consuming: it was trained for 40 h on two Nvidia V100 graphics cards with a batch size of 256. The aesthetic condition GAN proposed by Zhang et al. requires training on datasets with both semantic and aesthetic labels, and there are no large-scale datasets that meet these requirements. Therefore, the semantic quality of the image generated by GAN under the existing aesthetic condition is greatly reduced, and its aesthetic improvement is also rather limited. In other words, the existing global aesthetic image generation algorithm cannot make use of the high-quality construction ability of the existing generated model, and it is difficult to learn and generate aesthetic images in small-scale datasets.

2.4. Databases

There are many datasets used for image aesthetic research, and different datasets can be studied based on different aesthetic tasks. ImageNet [18] is a large visualization database for visual object recognition software research, which has over 14,000,000 images, over 20,000 categories, and more than 1,000,000 images with explicit category annotation and object position annotation. This dataset remains one of the most commonly used datasets for image classification, detection, and localization in the deep learning field. CelebA [19] is a large-scale face attributes dataset consisting of 200,000 celebrity images and every image has 40 attribute annotations. CelebA and its associated CelebA-HQ [20], CelebAMask-HQ [21], and CelebA-Spoof [21] are all widely used in the generation and manipulation of face images. LSUN [22] is a scene-understanding image dataset, which mainly contains 10 scene categories such as bedroom, living room, church, and 20 object categories such as birds, cats, and buses, with a total of about 1 million labeled images. AVA [23] is a dataset for aesthetic quality evaluation that contains 250,000 images, and every image has a series of ratings as well as 60 classes of semantic-level labels. The dataset also contains 14 categories of photographic styles such as complementary colors, duotones, and light on white, etc. Our study is based on the AVA dataset and the GAN decoupling representation for the generation of global aesthetic images.

3. Method

Image generation is the process of mapping the hidden space Z to the image space through a pre-trained generative model G, namely $G : Z \rightarrow X$. The research on the global aesthetic image generation algorithm focuses on how to map the latent space to an image space with global aesthetics.

This paper proposes a global aesthetic image generation method based on the disentangled representation of the generated adversarial network. This method uses the ability of the high construction of the existing generative model to the image content, learns the disentangled representation of global beauty from the latent space, and directly edits the global aesthetic feeling of the latent space to add global aesthetic features to the generated

images, so as to realize the generation of images with both high-quality image content and global aesthetic feeling.

The algorithm mainly includes two steps. The first step is to learn the decoupling representation of global beauty from the hidden space of GAN to obtain the global aesthetic decision surface as separated as possible. The second step uses the global aesthetic decision to aesthetically edit the hidden space of GAN and send it to the generative model to generate the global aesthetic image. The first step consists of three modules: generative model, prediction model, and SVM classifier, so the algorithm is generally composed of four parts.

The overall framework of the algorithm is shown in Figure 2. The algorithm first performs image generation through the StyleGAN model and scores the generated images through the NIMA prediction model. Next, the global aesthetic SVM decision face is learned from the Z and W latent spaces of the generative model, based on the generated image and its prediction score. Finally, the resulting decision is aesthetically edited in the latent space of the image to obtain the image with global beauty.

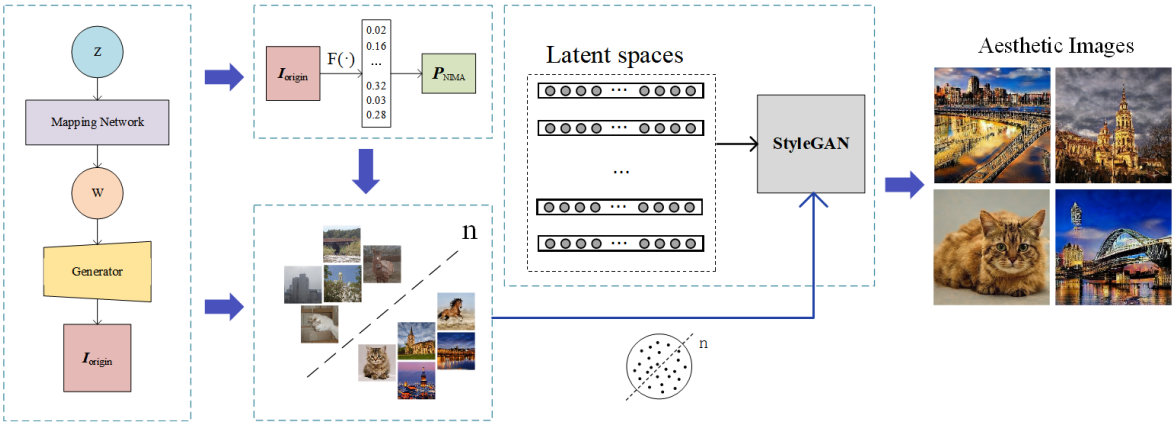

Figure 2. The generation process of global aesthetic image.

3.1. The Model of Generating and the Image Generation

In this section, the generation model used for the algorithm was introduced in detail, and the specific steps of image generation were described. All the generative models used in this experiment are based on the generative models of StyleGAN pre-trained on the LSUN [22] dataset. StyleGAN learned a more decoupled latent space W based on the input Z latent space of the conventional GANs and fed the w latent space into each convolutional layer of the StyleGAN generating network for different transformations. And it is different from other GAN models which send the latent space into the first layer of the network. Thus, StyleGAN makes the W latent space even more decoupled. And the structure of the StyleGAN network is shown in Figure 3.

Aesthetic style image generation algorithms based on GAN decoupling representation learning first use the generative model to generate large-scale image samples as sample sets for prediction classification, which are used to screen images that are aesthetically different as much as possible as positive and negative samples for style decision surface training. This is due to the fact that the properties of the images randomly generated by the generated model are uncontrollable, and the proportion of negative samples (i.e., images with low prediction style probability) in the generated images is larger. Therefore, in order to obtain a sufficient number of positive samples (i.e., images with high prediction style probability), it is necessary to generate a sample set as large as possible. This paper considers the experimental needs and hardware support, and finally the initial sample set number is set to 500,000. This experiment produces global aesthetic image generation of 11 different objects based on 11 generative models provided by StyleGAN. The 11 generation models are three outdoor scenes of church, bridge, and tower, four indoor scenes models of bedroom,

apartment, classroom, and conference room, and three specific object generation models of cat, car, and horse. Perform the following operations on 11 generative models, respectively.

1. Randomly generate 500,000 random latent spaces z to form the latent space Z and save Z;
2. The mapping network feeding Z into StyleGAN gives the latent space W and saves W;
3. Input W into the StyleGAN generator to get 500,000 images, saved as I_{origin};
4. Predict scores for I_{origin} separately to obtain P_{NIMA}.

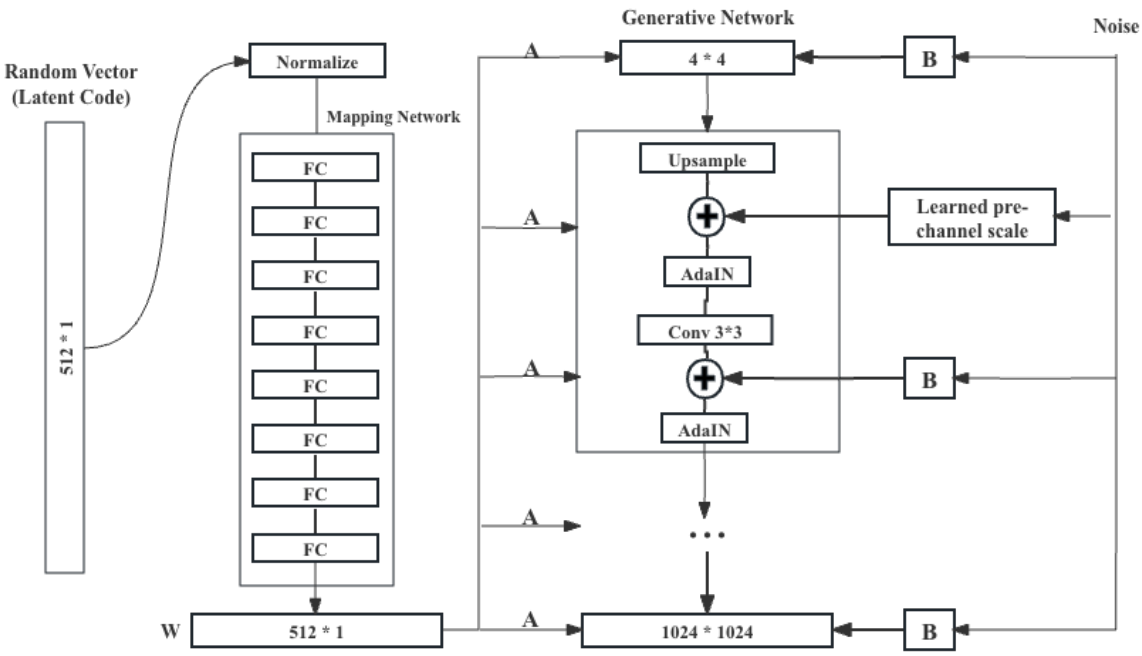

Figure 3. The structure diagram of StyleGAN.

3.2. The Predictive Model and the Global Aesthetic Score

This paper uses NIMA as the model for predicting the global aesthetic quality of images. NIMA is a convolutional neural network-based model for predicting aesthetic image distribution, proposed by Google Research. It achieves an aesthetic prediction accuracy of 80.6% on the AVA dataset. Compared to other aesthetic evaluation models with similar accuracy, NIMA has a simpler structure and faster training speed. In this paper, we implemented the NIMA algorithm and trained the model on the AVA aesthetic dataset. The training process and results of NIMA are presented below.

3.2.1. Dataset and Preprocessing

NIMA was trained based on the AVA aesthetic dataset. The AVA aesthetic dataset contains more than 250,000 images, each voting from 78–594 people from 1 to 10 points, with a label distribution in the form of $D = \{(x^1, d^1), (x^2, d^2), \ldots, (x^n, d^n)\}$ which represents the image, representing the voting distribution of the image, dimension [1, 10]. In order to train the NIMA model, two pre-processing steps were performed on the AVA dataset in the experiment. In the first step, to standardize the calculation method, the scoring distribution of each image was transformed into a probability distribution as the ground truth. In the second step, the AVA dataset was split into training and testing sets in an 8:2 ratio for iterative training. The method for calculating the probability distribution is shown in Formula (1).

$$y_i = \frac{v_i}{\sum_{k=1}^{c} v_k} \tag{1}$$

where c is 10, which represents the highest score for rating. The value of k corresponds to a certain score, v_k represents the number of voters for this score, $\sum_{k=1}^{c} v_k$ represents the total number of people scoring the image, and the probability of obtaining score i, denoted as y_i, is calculated by dividing the number of votes corresponding to that score by the total number of voters.

3.2.2. The Model Structure

The overall structure of the NIMA model is based on the traditional image classification network (VGG16, MobileNet, Inception-v2), the last layer of the original classification network is replaced with a fully connected layer, and the output dimension is set to 10, and the prediction probability of ten scores is obtained by a softmax activation function. Ground truth distribution of human ratings of a given image can be expressed as an empirical probability mass function $p = [p_{s_1}, ..., p_{s_N}]$ with $s_1 \leq s_i \leq s_N$, where s_i denotes the ith score bucket, and N denotes the total number of score buckets. In the AVA dataset, $N = 10$, $s_1 = 1$, and $s_N = 10$. And the prediction probability sums up to 1, as in Formula (2). And we can qualitatively compare images by their mean and standard deviation of scores, as in Formula (3) and (4).

$$\sum_{i=1}^{N} p_{s_i} = 1 \tag{2}$$

$$\mu = \sum_{i=1}^{N} s_i \times p_{s_i} \tag{3}$$

$$\sigma = (\sum_{i=1}^{N} (s_i - \mu)^2 \times p_{s_i})^{\frac{1}{2}} \tag{4}$$

In this paper, we trained NIMA with VGG16 as the base model, and the weights of VGG16 were initialized by the pre-trained weights on the ImageNet dataset [18] and the added fully connected layer weights. NIMA calculates the loss between the predicted value and the true value using the Earth Mover's Distance as a loss function, and iteratively updates the parameters by back-propagation against the loss. The *EMD* is defined as the minimum cost to move the mass of one distribution to another. Given the ground truth and estimated probability mass functions p and $\hat{p}$, with N ordered classes of distance $\|s_i - s_j\|_r$, the normalized Earth Mover's Distance can be expressed as

$$EMD(p, \hat{p}) = (\frac{1}{N} \sum_{k=1}^{N} \mid CDF_p(k) - CDF_{\hat{p}}(k) \mid^r)^{\frac{1}{r}} \tag{5}$$

where $CDF_p(k)$ is the cumulative distribution function as $\sum_{i=1}^{k} p_{s_i}$. And it is worth noting that this closed-form solution requires both distributions to have equal mass as $\sum_{i=1}^{N} p_{s_i} = \sum_{i=1}^{N} \hat{p}_{s_i} = 1$. The model will perform aesthetic prediction scores for the images generated by the generative model, as well as to evaluate the generated aesthetic image, as an evaluation indicator of the effect of the aesthetic image generation. Since the original output of NIMA is a probability distribution of 110 points of the image, by calculating the mean of the distribution as the aesthetic score of the image, we calculate the Formula (6), wherein score represents the aesthetic score of the image, the variation interval of i is [1, 10] corresponds to aesthetic score 1–10 points, and pi corresponds to the prediction probability of each score.

$$score = \sum_{i=1}^{n=10} i * p_i \tag{6}$$

3.2.3. The Global Aesthetic Decision Boundary

Support vector machine (SVM) is a supervised learning model for analyzing data in classification and regression analysis, which is a classical binary classification machine

learning algorithm. Given the data and labels, SVM can learn a hyperplane called the decision boundary that separates the attributes as much as possible. A linear problem is the initial core problem that SVM aims to solve. For a given dataset with corresponding labels, $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, $x_i \in R^n$ is the target data of the learning decision surface, and y_i and x_i are the one-to-one label value. SVM learns a hyperplane, denoted as $w^T x + b = 0$, where (w) is a weight vector, (x) is an input vector, and (b) is a bias term. The hyperplane $w^T x + b = 1$ minimizes the distance between positive samples, while the hyperplane $w^T x + b = -1$ minimizes the distance between negative samples. The problem solved by SVM is to find a hyperplane that maximizes the margin between the two sides while minimizing the probability of misclassification. The formula is shown below:

$$\gamma = \frac{2}{\|\omega\|} \tag{7}$$

$$\begin{aligned} &\arg \min_{\omega,b} \frac{1}{2} \|\omega\|^2 \\ &s.t. y_i(\omega^\top x_i + b) \geq 1, i = 1, 2, ..., m. \end{aligned} \tag{8}$$

where ω and b are the parameters of the hyperplane, and γ is the distance between hyperplane $w^T x + b = 1$ and hyperplane $w^T x + b = -1$.

This algorithm utilizes SVM to learn decoupled representations of aesthetic styles from the hidden space of the generative model. Aesthetic SVM decision faces are, respectively, learned from the Z and W latent spaces of the generative model that is based on I_{origin} and P_{NIMA}. The 2000 sheets with the highest NIMA score and the 2000 sheets with the lowest NIMA score were selected as the aesthetic decision faces corresponding to the training of positive and negative samples. Examples of the positive and negative samples of the partially generated model are shown in Figure 4.

Finally, a total of 22 aesthetic decision faces were trained in Z and W of 11 generated models, and the classification accuracy of decision faces is shown in Table 1. The average accuracy of all decision faces based on W latent space training is higher than that on Z latent space, and the average accuracy improves by 9.3%. The accuracy pairs, such as seen in Figure 5. The average accuracy of the decision surface based on W latent space training reached 0.997.

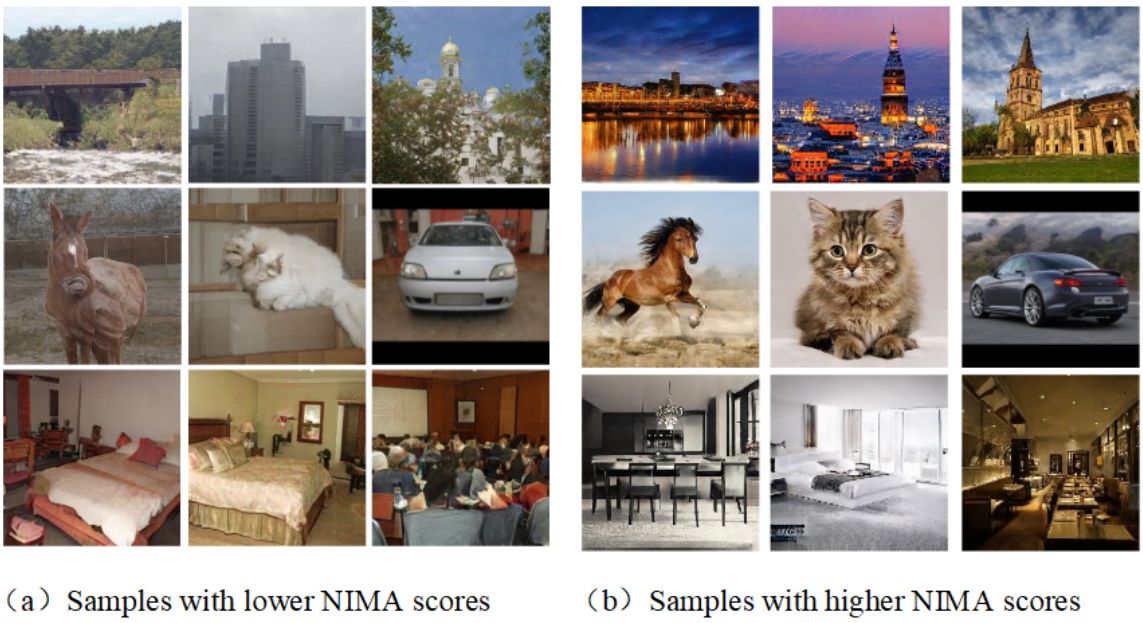

Figure 4. The example of positive and negative samples for SVM training.

Table 1. Generation model NIMA decision surface accuracy.

	Church	Bridge	Tower	Bedroom	Apartment	Classroom	Boardroom	Cat	Car	Horse
Z	0.866	0.95	0.886	0.943	0.95	0.956	0.903	0.92	0.91	0.83
W	0.999	0.998	0.999	0.998	0.999	1.000	0.999	0.998	0.99	0.99
Improve	15%	5%	13%	6%	5%	5%	10%	8%	8%	19%

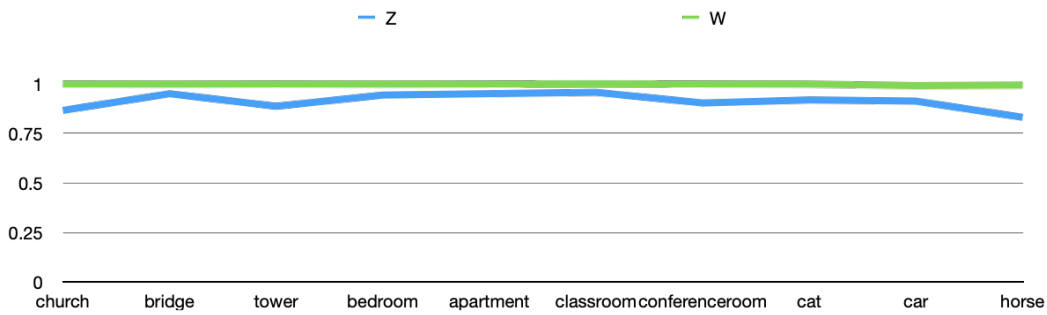

Figure 5. The comparison of SVM classification accuracy of different hidden spaces of global aesthetic decision surface.

3.2.4. Aesthetic Editor

The image generation experiments presented in this paper were all completed based on StyleGAN and StyleGAN2. The disentanglement of StyleGAN is the biggest feature that distinguishes this model from other generative models. It maps the Z latent space of traditional GANs to the more decoupled W latent space through a mapping network, so it is more conducive to the learning of various semantic disentangled representations. Based on this, the global aesthetic generation experiment of this paper edited the aesthetic style of the latent space of Z and W, respectively, in which any latent space in Z and W corresponds one-to-one. That is, any $z_i \in Z$ and the corresponding $w_i \in W$ will obtain the same original image through the generative model.

Using the obtained global aesthetic decision surface, three different editing experiments are the linear editing of the original latent space Z, linear editing of the more decoupled W latent space, and editing of the input global aesthetic representation to different levels of the generative network.

For the linear editing of Z, the Z is linear to the direction of the global aesthetic, such as seen in Formula (9), where n is the global aesthetic decision surface, the linear editing step is in the direction of n, increased to n such that $z' = z_i + 3n$ is the direction of the global aesthetic improvement promoted in the three steps.The edited input mapping network becomes a total of 14 layers of the StyleGAN network with global aesthetic decoupling representation and input after different transformations, and finally obtains the image with a global aesthetic feeling. This process can be expressed as Formula (10), in which G is the generation model and y is the global beauty of the original latent space. The linear editing process for W is to regularize the random vector z to obtain the latent vector w at first, through a nonlinear transformation network, and then the linear editing for w is the same as the above process.

$$z' = z_i + \lambda n \tag{9}$$

$$y = G(z + \lambda n) \tag{10}$$

4. Experiment and Result Analysis

This chapter mainly discusses the experimental setup and procedure, presents and analyzes the experimental results corresponding to different latent spaces, as well as provides an analysis of the results of global aesthetic image generation.

4.1. Experiments Setup

We introduce the experimental setting of the proposed algorithm. For the generative model G described in Section 3.1, the 11 generating models are 11 different objects. During image generation, first, the randomly generated 5,000,001,512 dimensional random latent space z forms the latent space Z, and then we input Z into the mapping network of StyleGAN to obtain the latent space W, which was input into the generator to obtain 500,000 images.

To train the predictive model, at first, the input image was first trimmed to the image size of 256×256, and then the 224×224-sized images were extracted with the random cropping method, thus reducing the speed of model overfitting. NIMA is trained using the loss of the validation set as a constraint on the number of training iterations, setting the stop iteration parameter that stops training when the loss on the training set exceeds that number.

This experiment was trained on a single NVIDIA 2080Ti with the number of iterations set to 100, batch size set to 64, convolution layer learning rate set to 0.005, and fully connected layer learning rate set to 0.0005. After each iteration, tests were conducted on the validation set, and the EMD loss and prediction accuracy were calculated. The EMD loss exceeded ten iterations and did not decrease when the training stopped.

In the learning stage of global aesthetic decision faces, we selected the 2000 highest NIMA scores and 2000 lowest scores as positive and negative samples in Z and W of the 11 generated models to obtain a total of 22 aesthetic decision faces. Finally, we also performed the aesthetic editing of Z and W separately to obtain a global aesthetic image. The experimental procedure is shown in Figure 6

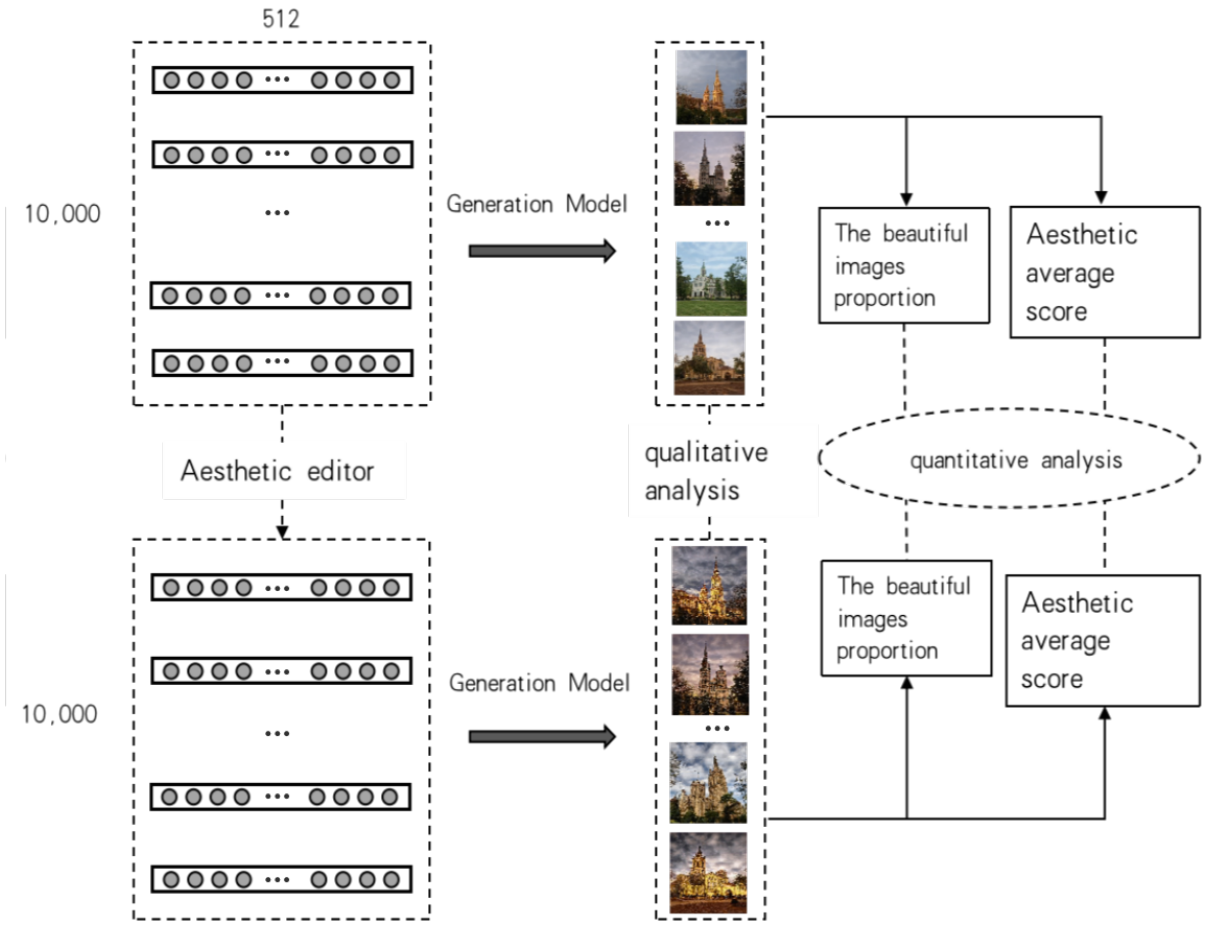

Figure 6. The procedure of experiment.

4.2. Quantitative Analysis of Experimental Results

In this section, we first conducted a quantitative analysis by comparing the image scores generated by editing the aesthetics on different latent spaces (W and Z) with the scores of the original images. And the experiments involved quantifying the average aesthetic scores and aesthetic image rates for both the original generated images and the aesthetically edited images using the NIMA (Neural Image Assessment) model. The corresponding improvement rates were also calculated. Furthermore, we used line graphs to visually demonstrate the relationship between image scores, image semantics, and the degree of aesthetic editing in the W latent space. Additionally, we provided evidence of the effectiveness of aesthetic generation through a histogram depicting the changes in the rate of aesthetically pleasing images.

The experiment used the NIMA aesthetic evaluation model as the aesthetic quality evaluation index of the generated images, predicting the 10,000 test images. The aesthetic score of the generated images after *Z* editing, the aesthetic score of the generated images after *W* editing, and the aesthetic average score of 10000 images was calculated for comparison. Table 2 shows the three scores for each generated model. Looking at the data in the table, we can see that the aesthetic scores of the edited images were improved, whether through the aesthetic editing based on the *Z* or *W* implicit space. In contrast, the image aesthetic score improved more after latent spatial aesthetics editing based on *W*.

Table 2. The average score of the aesthetic grading.

	Church	Bridge	Tower	Bedroom	Apartment	Cat	Car	Horse
Initial	4.8189	4.916	4.9875	4.3932	4.5333	4.4196	5.1415	4.7391
Z	5.4915	5.4606	5.3251	5.4279	5.5754	5.6024	5.9814	5.3482
W	5.6506	6.0398	6.2284	5.6086	5.6986	5.7370	5.6924	5.3797

Figure 7 shows the aesthetic mean score contrast mixed histogram, where the yellow discount represents the initial NIMA score, the orange bar represents the aesthetic average score of the image generated based on *Z*-editing, and the blue bar represents the aesthetic average score of the image generated based on *W*-editing. It can be seen that *Z* greatly improves the aesthetic quality of the images generated by *W* editing compared with the original images. Moreover, from the histogram, it can be observed that the editing effect on *W* is almost entirely better than that on *Z*, and only the aesthetic image generation of "car" is slightly less satisfactory than the editing effect on *Z* hidden space, which is consistent with subjective perception.

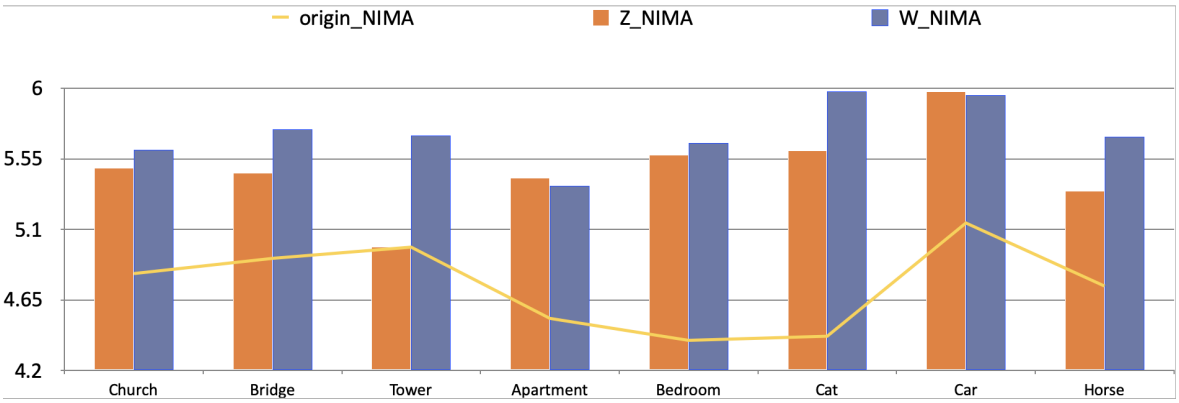

Figure 7. The mixed histograms of the NIMA scores contrast.

The NIMA average aesthetic score for the original generated and aesthetic generated images are calculated in the experiment. The beauty rate was quantified, and the corresponding improvement rate was calculated. Detailed data are presented in Table 3; the

algorithm achieved excellent results in improving the aesthetic average score. Comparing the aesthetic average score with the original, the first generated image was increased by 22.26%. The beauty rate also achieved an almost perfect effect, increasing the average beauty rate by 71%, and bridges, towers, and cars finally even reached 100%.

Table 3. Improvement rate of aesthetic indicators.

	Average Aesthetic Score			Two Classification of "Good-Looking" Proportion		
	I_{origin}	$I_{aesthetic}$	Increase Rate	I_{origin}	$I_{aesthetic}$	Increase Rate
church	4.8190	5.6506	17.3%	0.212	0.9962	78.42%
bridge	4.9159	6.0398	22.9%	0.3796	1.0000	62.04%
tower	4.9875	6.2284	24.9%	0.4535	1.0000	54.65%
bedroom	4.3932	5.6086	27.7%	0.0601	0.9976	93.75%
apartment	4.5336	5.6986	25.7%	0.1239	0.9955	87.16%
cat	4.4196	5.737	29.8%	0.0610	0.9986	93.76%
car	5.1415	5.9814	16.3%	0.7863	1.0000	21.37%
horse	4.7393	5.3797	13.5%	0.1529	0.9397	78.68%
average value	4.7437	5.7905	22.26%	0.3097	0.9910	71.23%

Eight generating models were edited 10 times, and the aesthetic average score of each yuan editing image was recorded. Figure 8 shows from the original image of 10 yuan editing images in the process of aesthetic score changes. It can be clearly seen from the eight generating models that, with each yuan edit, the latent space is moved in the direction of aesthetic quality, as the aesthetic average score rises.Horses, churches, cats, and cars maintain the same increasing rate almost always, while the generative models gradually decrease. After eight yuan edits, the aesthetic average score of all the generated images tends to flatten out and fluctuate slightly.

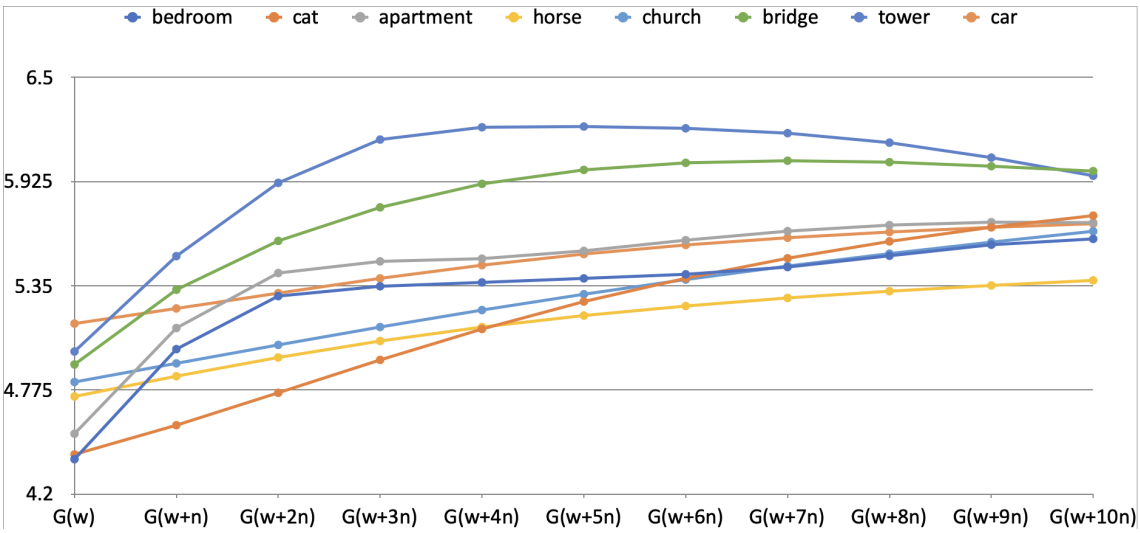

Figure 8. The line plot of NIMA mean score change.

Figure 9 shows the eight-generation model binary classification as "good" image (beauty rate) changes. It can be clearly seen from the figure that the beauty rate of the original image is uneven; the bedroom and apartment indoor scenes' beauty rate is relatively low, at less than 1.5%, and the car's beauty rate is far greater than in other generation models, reaching 78%. However, after 9 to 10 steps of aesthetic editing, almost all of the eight generative models generated a 100% beauty rate. This illustrates the effectiveness of our aesthetic image production.

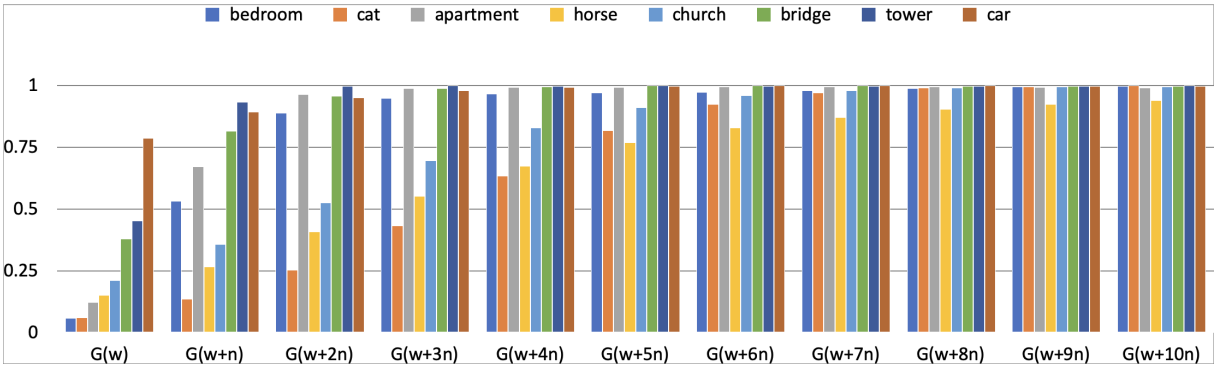

Figure 9. The change histogram of the rate in beautifying images.

4.3. Qualitative Analysis of Experimental Results

In this section, we conducted a qualitative analysis of the experimental results to demonstrate the effectiveness of our global aesthetic image generation algorithm. We compared the images generated from different latent spaces with the original images, as well as compared the original images with aesthetically edited images of different semantics. These comparisons were conducted to showcase the effectiveness of our algorithm in generating aesthetic images.

To better capture the aesthetic information in the generated aesthetic images, we generated 10,000 images and used them to verify the effectiveness of the proposed aesthetic image generation algorithm. Based on the 22 global aesthetic decision faces obtained in Section 3.2.3, the Z and W latent spaces of the 10,000 generated images were aesthetically edited, respectively, and the corresponding global aesthetic images were obtained. After observation, the generated images obtained by editing on different latent spaces had significant visual differences. Therefore, the qualitative and quantitative analysis of the global aesthetic generation effect on the Z and W latent space was conducted first. Figure 10 shows an example of the change in the image during the linear aesthetic editing of Z and W of StyleGAN's generated church.

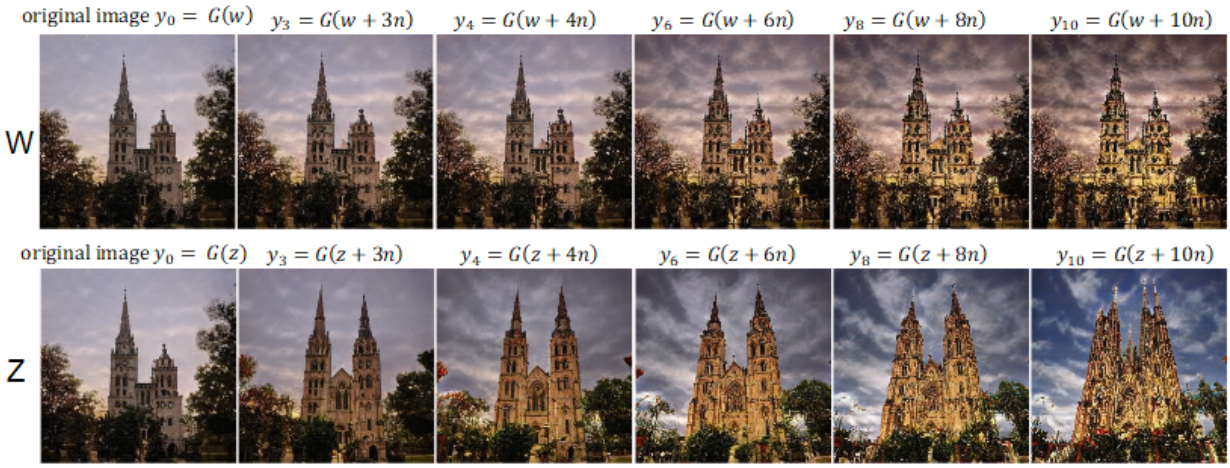

Figure 10. The example of the church with Z and W editing processes.

Where G represents the generative model, w represents a latent space in W space, z represents a latent space in Z space, $y = G(W)$ represents the image obtained from the original w input G, $y_3 = G(w + 3n)$ pushes the w vector in three steps in the direction of aesthetic evaluation promotion, and so on. The image aesthetic effects generated in the W and Z latent spaces are rather different and, in the same latent space, the number of aesthetic steps is different, and the effect also presents a big gap. The larger the number of steps, the stronger the aesthetic degree.

Figure 11 shows the results of w (line 1), 3, 4, 6, 8, 10, and z (line 2) in the direction of aesthetic evaluation, respectively. It can be seen that W-based aesthetic editing has a much less semantic impact on the original generated images than Z-based aesthetic editing. The editing of the W hidden space causes little semantic deformation of the image, and the overall structure is almost consistent with the original image. The final image generated by Z is rather different from the original image, but compared with the original image, both achieved the aesthetic quality improvement effect visible to the naked eye. At the same time, red impurities can be observed in Z in the gradual editing process. After observing many examples of Z, the probability of impurities in editing is much higher than for W. The following figure shows some examples of W and Z generation. In this figure, based on the generated image of the W latent space, the overall semantic structure of the cat is almost unchanged, but visually, the image quality is better. Although the image of the cat based on Z space is still a cat, the overall structure of the image has changed greatly, and the body of the cat has disappeared. Thus, the aesthetic image based on the W latent space is more in line with our requirements.

And we analyze the global aesthetic image generation results based on editing results on W latent space. Figure 12 shows a partial sample of the aesthetic image generated based on the algorithm proposed in this study. Most of the aesthetic improvement makes the details of the image richer, the color becomes more in line with the public aesthetics, or it eliminates part of the noise. There are different degrees of aesthetic improvement visible to the naked eye. In the experiment, the quantitative indicators of the original and aesthetic generated images are analyzed to prove the actual improvement effect of aesthetic quality, namely the effectiveness of the aesthetic image generation algorithm.

Figure 11. The comparison of the aesthetic images generated by W and Z.

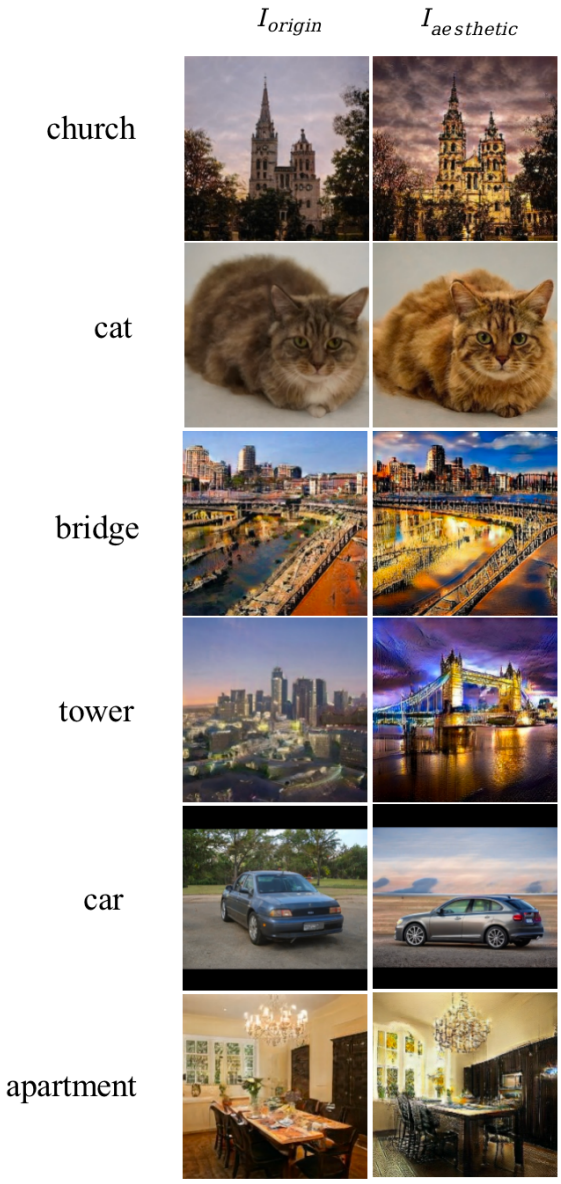

Figure 12. The example of aesthetic image generation results for global aesthetic improvement.

5. Conclusions

In this study, we used the existing StyleGAN generation model to propose a global aesthetic image generation algorithm based on GAN's interpretability to generate both semantic and aesthetic images. The algorithm effectively improved the aesthetic average score of the generated model and the rate of beauty generation. We first trained a global aesthetic prediction model, and then used the model to score and screen the initial image space generated by StyleGAN. This score was used as a label to learn the linear global aesthetic decision surface from the latent space through the support vector machine classifier. The results show that the decision surface learned from the W latent space had a classification accuracy of 99% or above. Based on this decision, we again generated a certain number of random latent spaces and global aesthetic editing to obtain the aesthetic and semantic generation image. The experiments showed that our method effectively improved the generation model image aesthetic average and developed a generation model with a 99% probability of the aesthetic binary classification task being defined as a "beautiful" image. In our approach, we utilized an SVM classifier to learn a linear aesthetic decision boundary from the latent space of the generative model. We also performed linear editing

on the latent space. Although this approach has yielded promising results in terms of aesthetic generation, upon observation, we still notice that the learned aesthetic representation is not fully decoupled, leading to semantic distortions in certain scenarios. Aesthetic perception in images cannot be adequately represented by a linear decision boundary alone. Aesthetics should be modeled and treated as a higher-dimensional problem. Thus, in future research, we will consider learning nonlinear aesthetic representations from the latent space to achieve a better decoupling between aesthetics and content, thereby enhancing the controllability of aesthetic image generation.

Author Contributions: H.Z., L.Z. and M.W. all contributed to the conception of the work. H.Z. conceived the model and M.W. and L.Z. conducted the experiments. H.Z., L.Z., Y.W. and Y.L. analyzed the results. M.W. wrote the main manuscript. All authors agree both to be personally accountable for their own contribution and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by National Natural Science Foundation of Zhejiang Province (LQ19F020008), the National Key Research and Development Program of China (No. 2017YFE0118200), and the National Natural Science Foundation of China (No. 61471150).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Only publicly available data were used in this article.

Acknowledgments: Thanks for support and assistance from Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province. Thanks for support and assistance from Key Laboratory of Network Multimedia Technology of Zhejiang Province.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jin, X.; Zhou, B.; Zou, D.; Li, X.; Sun, H.; Wu, L. Image aesthetic quality assessment: A survey. *Sci. Technol. Rev.* **2018**, *36*, 10.
2. Kao, Y.; Wang, C.; Huang, K. Visual aesthetic quality assessment with a regression model. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP),Quebec City, QC, Canada, 1–27 September 2015 ; IEEE: Quebec City, QC, Canada, 2015; pp. 1583–1587.
3. Jin, B.; Segovia, M.V.O.; Süsstrunk, S. Image aesthetic predictors based on weighted CNNs. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP),lPhoenix, AZ, USA, 25–28 September 2016; IEEE: Phoenix, AZ, USA, 2016; pp. 2291–2295.
4. Zeng, H.; Zhang, L.; Bovik, A.C. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv* **2017**, arXiv:1708.08190.
5. Zhang, X.; Gao, X.; Lu, W.; He, L.; Li, J. Beyond vision: A multimodal recurrent attention convolutional neural network for unified image aesthetic prediction tasks. *IEEE Trans. Multimed.* **2021**, *23*, 611–623. [CrossRef]
6. Talebi, H.; Milanfar, P. Nima: Neural image assessment. *IEEE Trans. Image Process.* **2018**, *27*, 3998–4011. [CrossRef] [PubMed]
7. Zaltron, N.; Zurlo, L.; Risi, S. Cg-gan: An interactive evolutionary gan-based approach for facial composite generation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; AAAI Press: Palo Alto, CA, USA, 2020; pp. 2544–2551.
8. Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 2642–2651.
9. Van, T.P.; lNguyen, T.M.; Tran, N.N.; Nguyen, H.V.; Doan, L.B.; Dao, H.Q.; Minh, T.T. Interpreting the latent space of generative adversarial networks using supervised learning. In Proceedings of the 2020 International Conference on Advanced Computing and Applications (ACOMP),l Quy Nhon, Vietnam, 25–27 November 2020; pp. 49–54.
10. Lee, W.; Kim, D.; Hong, S.; Lee, H. High-fidelity synthesis with disentangled representation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 157–174.
11. Shen, Y.; Yang, C.; Tang, X.; Zhou, B. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2004–2018. [CrossRef] [PubMed]
12. 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, Long Beach, CA, USA, 15–20 June 2019; pp. 142–149.

13. Li, H.; Wang, D.; Zhang, J.; Li, Z.; Ma, T. Image super-resolution reconstruction based on multi-scale dual-attention. *Connect. Sci.* **2023** , 1–19 [CrossRef]
14. Li, H.-A.; Hu, L.; Zhang, J. Irregular mask image inpainting based on progressive generative adversarial networks. *Imaging Sci. J.* **2023**, *71*, 299–312. [CrossRef]
15. Brock, A.; Donahue, J.; Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv* **2018**, arXiv:1809.11096.
16. Murray, N. Pfagan: An aesthetics-conditional gan for generating photographic fine art. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3333–3341.
17. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
18. Deng, J.; Dong, W.; Socher, R.; Li, L. J.; Li, F. F. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
19. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 3730–3738.
20. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
21. Lee, C.-H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitione, Seattle, WA, USA, 14–19 June 2020; pp. 5549–5558.
22. Yu, F.; Zhang, Y.; Song, S.; Seff, A.; Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv* **2015**, arXiv:1506.03365. [CrossRef]
23. Murray, N.; Marchesotti, L.; Perronnin, F. Ava: A large-scale database for aesthetic visual analysis. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, lDaejeon, Republic of Korea, 5–6 November 2012; pp. 2408–2415.