International Journal of Web-Based Learning and Teaching Technologies
Volume 18 • Issue 2

Application of Multimedia Data Feature Extraction Technology in Teaching Classical Oil Painting

Zhuo Chen, Harbin University of Commerce, China
Jianmiao Li, Shaoxing University, China*

ABSTRACT

The cross-modal oil painting image generated by traditional methods makes it easy to miss the important information of the target part, and the generated image lacks realism. This paper combines the feature extraction technology of multimedia data with the generation confrontation network in deep learning, puts forward a generation model of classic oil painting, and applies it to university teaching. Firstly, the key frame extraction algorithm is used to extract the key frames in the video, and the channel attention network is introduced into the pre-trained ResNet-50 network to extract the static features of 2D images in short oil painting videos. Then, the depth feature mapping is carried out in the time dimension by using the double-stream I3D network, and the feature representation is enhanced by combining static and dynamic features. Finally, the high-dimensional features in the depth space are mapped to the two-dimensional space by using the opposition generation network to generate classic oil painting pictures.

KEYWORDS

Dynamic Features, Fight Against the Network, Generation of Classical Oil Paintings, Multimedia Data Feature Extraction, Static Characteristics

INTRODUCTION

Image, as the most direct expression of art, penetrates life, and in the field of deep learning-based data processing, many problems can be seen as image transformation tasks. The technical results of image conversion can be used both in the creation of new forms of contemporary artwork and in many other areas of data processing. In the last few years, as deep learning research has become more advanced, image transformation models and corresponding algorithms have been rapidly developed, in particular, the emergence of the generative adversarial network (GAN) (Yi, et al., 2019) and its improvement. GAN and image transformation technologies are also used for addressing real-world issues, such as model prediction and parameter identification, high-resolution reconstruction of images,

DOI: 10.4018/IJWLTT.333601 *Corresponding Author

image semantic segmentation, image restoration, image depth estimation, etc. These results show us that generative adversarial networks, as well as image transformation techniques, can be used to solve problems, such as data processing and image transformation in different domains (Cai, et al., 2021).

The creation of classical oil paintings, with the help of single still images, is time-consuming and laborious, both in terms of collecting images and labeling. A multimedia database is a database that stores and manages a large number of multimedia objects, such as audio data, image data, video data, sequence data, and hypertext data. Thus, how to mine the effective data of multiple modalities, such as video, audio, and text contained in multimedia and extract the correlation features of the research work is one of the main research topics in the area of computer vision, currently. Multimedia data mining includes many aspects. An example is image feature mining, including similarity search feature extraction, data cube, and multidimensional mining, association mining and classification, and predictive analysis.

In multimedia data cube and multidimensional feature extraction method, multimedia data cube is a kind of data cube for storing multidimensional data that also implements abstract data structures for multi-dimensional integrated queries at different abstraction layers, which can well support online analytical processing operations and data mining of multiple knowledge at multiple levels (Tang, et al., 2020). The complexity of multimedia data makes the structure of the multimedia data cube more complex with each additional dimension, and its physical implementation is a greater challenge. In practical applications, concepts of higher abstraction levels are often used instead of precise color values to represent knowledge rules, and conceptual inheritance can be used to generalize the original attributes to simplify the multimedia data cube structure. In addition, for multi-valued attributes, such as multiple colors of an image, the ones with the highest number of pixels of that color can be selected to make the corresponding multimedia data cube greatly simplified, and the mechanism of online analysis and mining is used to make the system with multiple data mining methods (Adnan & Akbar, 2019).

To mine the association of multimedia objects, each image can be viewed as a thing, from which patterns of high frequency can be identified. But the mining of association rules for multimedia objects are different from transactional database mining. First, an image can contain multiple objects, each of which can have many features, such as color, texture, shape, position, keywords, etc., such that a lot of associations exist. In many cases, a feature of two images is the same at a certain resolution level but is different at a finer resolution level. So a resolution progressive refinement approach is needed. Patterns with high frequency can first be mined at a relatively coarse resolution, and then further finer resolutions can be mined for those images that pass the minimum support value. Multimedia objects usually have a spatial relationship, such as up and down, left and right, front and back, etc., and these features are of greater significance for mining the association and relevance of objects. The relationship between space and other colors, textures, shapes, etc. can form meaningful associations. Thus, data mining methods about spatial aspects are very important for multimedia mining. Other important concepts include, implementing processing techniques, such as slicing, dicing, drill-down, spin-up, etc., data mining methods for knowledge discovery, and data mining to discover relationships between media features, classification of images, and videos based on media features, etc., based on user requests for multimedia feature libraries. Interactive or automated knowledge mining can be implemented to discover implicit knowledge of interest to users.

This paper combines multimedia data feature extraction technology with the generation of confrontation network in deep learning, proposes a generation model of classic oil painting, and applies it to university teaching. First, the key frame extraction algorithm is used to extract the key frames in the video, and the channel attention network is introduced into the pre-trained ResNet-50 network to extract the static features of the 2D image in the short oil painting video. Then, with the help of dual-stream I3D network, the depth feature mapping is performed on the time dimension to enhance the feature representation by combining static and dynamic features. Finally, the antagonism generation network is used to map the high-dimensional features in the depth space to the two-dimensional space

to generate the classic oil painting pictures. This article aims to solve the problems of traditional methods in generating cross modal oil painting images, such as lack of realism and possible omission of important information about the target part. The research issue of this article is how to combine multimedia data feature extraction technology with generating confrontation networks to improve the generation of classic oil paintings.

RELATED WORKS

Recently, deep learning techniques are advancing at a rapid pace, using the deep learning and multimedia data feature extraction and other related techniques in the design of classical oil paintings, human faces, landscape paintings, and other works and applying them to teaching tasks. This practical application has gained the attention of a wide range of researchers. This paper gives an analysis of the current state of research and its results at home and abroad from two aspects: multimedia data feature extraction techniques and deep learning theory and generative adversarial networks.

Multimedia Data Feature Extraction

Usually, the multimedia data classification process consists of two processes: feature extraction and classification. The features obtained after feature extraction provide more accurate descriptions of feature characteristics and more detailed measurements, which play a vital role in the performance of the classifier.

In the early stage of research on feature extraction techniques for multimedia data, a series of classical methods for feature space dimensionality reduction have been proposed, such as independent component analysis (ICA) (Sompairac, et al., 2019), principal component analysis (PCA) (Hasan & Abdulazeez, 2021), linear discriminant analysis (LDA) (Wen, et al., 2018), etc. However, the methods based on linear transformations, and so on, do not solve the nonlinear problems that exist in multimedia data, until the proposal of stream learning. Stream shape learning is applied to multimedia data, aiming to obtain the intrinsic structure of nonlinearly distributed data and achieve nonlinear dimensionality reduction of features. Without dimensionality reduction, some algorithms use kernel functions to map the input sample space to a higher dimensional feature space to make the complex nonlinear data structure in the feature space become simple and reduce the complexity of the processing.

Current image acquisition sensors can acquire images with higher spatial resolution and provide richer detailed information. Using joint spatial-spectral features instead of single image features, the classification results obtained are better than those of arbitrary spectral or spatial features. Besides, sparse representation is also an area of research that has received much attention for several years, and the literature (Ling, et al., 2018) proposes a dictionary-based sparse representation method applied to multimedia data classification tasks. In the literature (Hossain, et al., 2022), a sparse stream shape learning method with multiple graph embeddings is designed based on the construction of sparse graph structures with full variational optimization and the introduction of spatial information. The literature (Qin, et al., 2019) introduces an extended morphological property cross-section algorithm based on sparse representation, which obtains the multi-scale spatial information of the image by the opening and closing operation, and achieves the classification by combining the spectral information. Compared with other null-spectrum classification methods, this method is easy to operate and has a good classification effect.

Deep Learning Theory and GAN

Generative adversarial networks (GAN) in the field of artificial intelligence is one of the most rapidly developing and relatively new research directions in deep learning and art integration research, and the training process is a special adversarial process of generating image networks, discriminating true and false image networks competing with each other, and finally determining the balance. Due to the unique advantages of generative adversarial networks in the fields of image processing and

image generation, many application studies have been conducted at home and abroad, and the results have been better applied.

Words and images are two important ways for people to experience the world. Correlating these two modalities is currently an important research topic in both fields. While rich and complex semantic features of a text can be extracted from images, synthesizing images directly from text is extremely complex, but the advent of GAN provides an unsupervised model to generate images. By extracting the important attributes in the text (e.g., space, the relationship of things, state of things, etc.), and then using the generator and discriminator in GAN to game each other's states, it makes it possible to embed attributes in images. As proposed in the literature (Yu, et al., 2020), the CycleGAN structural model defines a forward GAN network from source data to destination data and a reverse GAN network from destination data to source data (Waheed, et al., 2020), respectively, forming a ring-shaped network structure and introducing a cyclic consistency loss function. In the absence of image label pairing data, the method completed a qualitative analysis and the results showed its superiority (Kahng, et al., 2018). The supervised model-based training process requires a lot of images and labels to be paired with one-by-one training data, images from different domains are transformed, and the data volume is even more enormous. The literature (Yan, et al., 2019) proposes the coupled generative adversarial network (Cogan) model, which uses two coupled GAN networks sharing the weight constraints and learning the joint distribution of images in the two domains with an unsupervised manner, controlling and processing the image generation process in two different domains separately to achieve cross-domain image generation. In the literature, Nandhini Abirami, et al. (2021), deal with the problem that the image super-resolution reconstruction process generates pseudo-textures and the local information of the original image may not be fully utilized. The super-resolution reconstruction method is based on attention generating adversarial network, by using an attention recursive network and introducing a dense residual block structure to achieve the generator to extract the local features from the image and the discriminator to complete the image correction to accomplish the goal of image super-resolution reconstruction. For the problem that different image conversion tasks require their specific conversion methods with no universal method, the literature (Wang, et al., 2020) proposes a generalized image transformation solution based on conditional generative adversarial network (CGAN), which can effectively perform the following tasks, such as composing pictures from labeled images, reconstructing image objects from line drawings, and coloring pictures. This framework can be applied to achieve reasonable image conversion without manually designing the mapping function or without manually designing the loss function. For large categories, small samples, multiple styles, unknown languages, and other complex text, it is difficult to achieve automatic completion of the problem. The literature (Roy, et al., 2018) uses global and local consistency preserving generative adversarial networks (GLCGAN), in the case of handwritten text, without writing style constraints to build a second-level complementation system and discusses the problems encountered when different missing parts of the text are instantiated, and it has better results in unconstrained handwritten Chinese character completions.

Previous studies have shown the potential of using deep learning techniques for image transformation tasks and their applications in various fields, including art creation and education (Chen, 2021). In particular, the use of generative adversarial networks has been proven to be effective in generating high-quality images with realistic features (Gatys, 2016). Furthermore, in the context of art education, the integration of multimedia data feature extraction technology and generative models can provide students with a more immersive and interactive learning experience (Zhu, 2017). For instance, it allows students to explore different styles and techniques of classical oil painting through the creation of their own artworks using generated images as references (Karras, 2018). Therefore, based on the existing literature, we argue that the proposed framework of combining multimedia data feature extraction and generative adversarial networks can contribute to the development of innovative and effective teaching methods in the field of classical oil painting. Previous work has explored the application of deep learning and multimedia data feature extraction techniques in various

fields, including art and painting. However, there are still gaps in the literature that the proposed study aims to address. Specifically, the proposed generative model of classical oil painting combines key frame extraction, channel attention networks, dual-stream I3D networks, and adversarial generative networks to enhance feature representation and generate realistic oil painting images. To the best of our knowledge, such a comprehensive approach has not been explored in the context of classical oil painting. The proposed study aims to fill this gap and shed light on the potential of multimedia data feature extraction techniques and deep learning in teaching classical oil painting.

Based on the existing theoretical research results on artificial intelligence and painting, this paper constructs a basic theoretical model in the intersection of multiple fields, such as art and art and deep learning. It is also applied to explore feasible paths, research methods, and related theories for the integration of classical oil painting art and natural science. This applied research, which is an important issue in the development of discipline construction, has both originalities, pioneering, integration, and innovative value, as well as the interdisciplinary characteristics of complexity, comprehensiveness, and frontiers characterized by the intersection of arts and sciences. It provides a new way of thinking for effectively promoting the innovative development of emerging and interdisciplinary subjects.

METHODOLOGY

This paper is intended for designing an end-to-end deep learning system consisting of an encoder, decoder, and discriminator. The encoder can automatically embed information into the complex texture region of the carrier image, and the decoder can extract information from the carrier image. The encoder receives the original oil painting carrier image *Ic* with the shape $C \times H \times W$, and then generates the image *Is* that hides the secret information. *Ic* and *Is* are perceptually identical in appearance. *Is* is the image encoded by the encoder, and its pixel values are all floating-point numbers. Then, inputting *Is* to the rounding layer, the output is *Is*' with integer pixel values. The decoder receives *Is*' and extracts the information embedded in it to obtain *Next*. The discriminator *Eve* receives the image and determines whether it is cover or stego and implements end-to-end optimization based on the loss.

Multimedia Feature Encoder

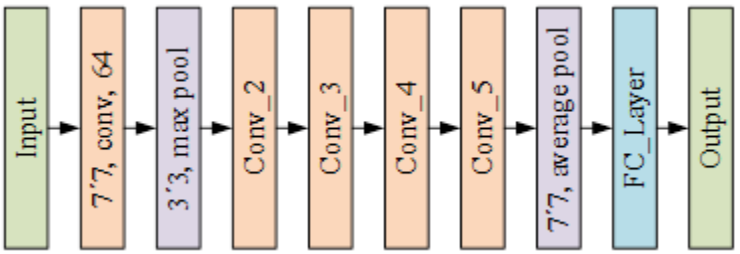
The difference between video data and picture data is that video is a multi-frame snapshot, and this makes video ideal for describing continuous action, utilizing real-time audio, and presenting a variety of different information composed of a visual and three-dimensional record of events. Compared with a single picture, video not only contains the space character, but also contains the time characteristics, and the characteristics of audio and movement. Since continuous frames in video convey a great deal of information, it is hard to identify the more salient content in context for accurate description. For this purpose, multiple modal features of fused video are used to perform the task of the textual representation of video content.

The squeeze and excitation (SE) (Rundo, et al., 2019) is used in the residual network ResNet50 to encode an image into deep space to extract the static features of videos, and a dual-stream expanded 3D convolutional network is used to extract dual-stream 3D features. The module incorporates the idea of dual-streaming into 3D convolution, which allows the network to better extract the spatio-temporal knowledge of the video, and features are captured with a fine-grained manner (Gu, et al., 2019).

ResNet and Channel Attention

Recently, tasks such as image semantic segmentation, target detection, and image classification have adopted ResNet-50 and ResNet-101 as backbone networks for feature encoder (Liu, et al., 2021). Therefore, this paper adopts ResNet-50 as the backbone network for feature coding, taking into account the experimental environment and recognition performance. The structure is shown in Figure 1. The ResNet-50 network contains 50 layers, where 49 layers are convolutional layers and 1 layer is fully connected. Because of its residual structure, it effectively solves the problems of gradient

Figure 1. ResNet-50 structure

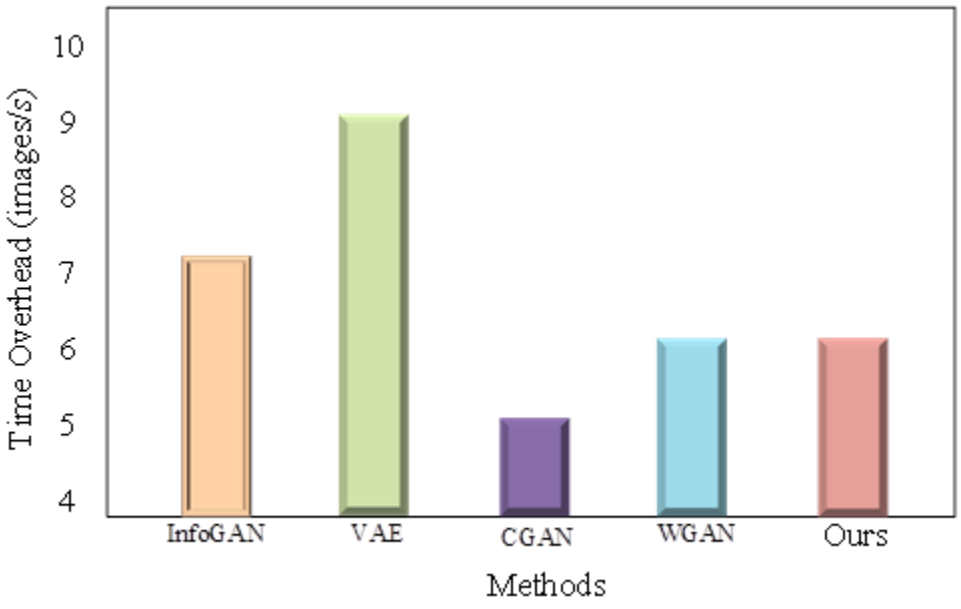


disappearance, gradient dispersion, and performance decay. First, after a convolutional layer of size 7 × 7 and a maximum pooling layer of 3 × 3, the deep features of the target object are extracted using multi-layer residual blocks to complete the feature encoding of the 2D image.

The attention mechanism, inspired by the human visual, learns the weights of parameters, and its core mission is to select key information that is more relevant to the current model objectives from the many pieces of information extracted. Frame-level feature extraction of the video extracts different information for each frame in different channels, thus, increasing the attention the channel can give greater weight to important features. In the SE module, the mechanism models channel dependencies on each other and adaptively recalibrates the channel-based feature vectors. Moreover, the global knowledge is employed to selectively emphasize important features and suppress redundant features. The structure of SE is shown in Figure 2.

Suppose given input *X*, and input data feature channel C', after convolution and pooling, the number of output data feature channels is C. The SE module then rescales the feature map U after the convolutional pooling process. In particular, the map U of the feature is compressed along the spatial dimension for the extrusion operation. The two-dimensional information in the feature channel

Figure 2. The framework of SE



is compressed into numbers Z_C, where Z_C is the global knowledge represented on the channel. Formally, the statistic Z_C is generated by reducing the spatial dimension (H×W) of the feature map *U*, therefore, the *c*-th element of *z* is calculated as in Equation (1).

$$Z_C = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

Excitation operations are performed to bring the advantage of the knowledge gathered, which is designed to completely capture channel dependencies. The Excitation operation is implemented using two fully connected (*FC*) (Pambala, et al., 2021) structures to reduce the complexity of the network structure and improve the overall generalization performance. The first *FC* layer acts as a dimensionality reduction, compressing the *C* channels into *c/r* channels, and the dimensionality reduction factor *r* is a hyperparameter. The second *FC* layer is applied to recover the original dimensionality of the feature map. A final weight coefficient *S* is obtained, which is calculated as shown in Equation (2).

$$S = F_{ex}(z, w) = \sigma(g(z, w)) = \sigma(w_2 \delta(w_1 z))) \tag{2}$$

where σ denotes the $\mathrm{sigmoid}$ function, δ denotes the relu function. $\mathrm{w}_1 \in R^{\frac{\mathrm{c}}{r} \times c}$, $\mathrm{w}_2 \in R^{c \times \frac{\mathrm{c}}{r}}$.

Lastly, the reweight operation is used, and the output of the weight is weighted with a channel-channel manner. The rescaling of the original features in the feature map channel dimension is used to obtain the final attentional features $\tilde{X}_C$. $\tilde{X}_C$ is calculated as shown in Equation (3).

$$\tilde{X}_C = F_{scale}(u_c, s_c) = s_c \bullet u_c \tag{3}$$

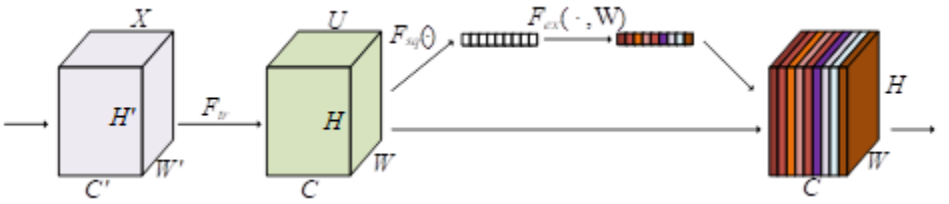
where F_{scale} is the channeled multiplication between attention weights s_c and feature maps u_c.

The SE eventually does attention or gating computing on the channel, and the attention mechanism makes the model focus on the most informative channel features and suppress the unimportant channel features.

Dual-Stream I3D Feature Extraction

To capture the temporal features in the video, the video 3D features are extracted using a dual-stream expanded 3D convolutional network structure I3D. Dual-stream I3D model is improved from the Inception-V1 (2D) model, whose base model structure is InflatedInception-V1. As shown in Figure 3(a), since dual streams can capture action information simply and effectively, this network structure introduces the idea of dual streams into 3D convolution to build I3D networks (Wang, et al., 2021). This consists of two 3D structures, one for receiving RGB information and the other for receiving optimized and smoothed optical flow information (Yang, et al., 2019). As shown in Figure 3. Using 2D structures, in which convolution kernels are repeatedly executed in the temporal dimension, to form 3D convolution kernels, where both the convolution kernel and pooling increase the overhead and the nonlinear structure in the model remains unchanged. Although 3D convolution is helpful for capturing the time-series features of videos, the idea of iteration is embedded in the optical flow algorithm. Adding optical flow to the 3D network structure can improve the model recognition accuracy.

Figure 3. 3D convolutional and dual-flow expanded 3D convolutional network structure



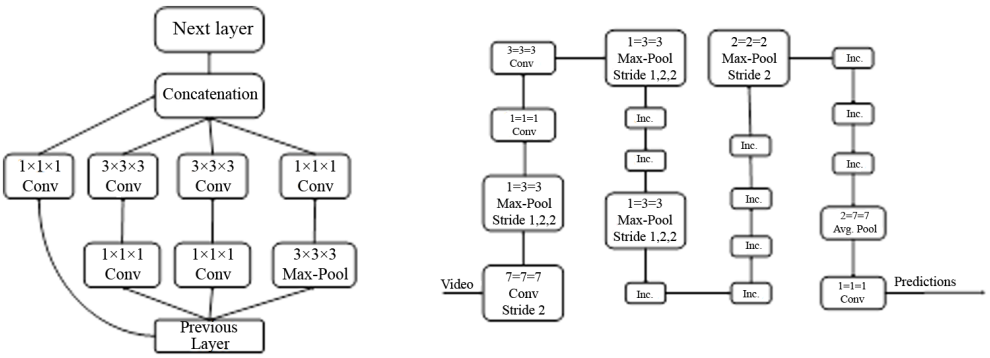
Generating Networks

The structure of the generator is shown in Figure 4. The whole network has a depth of 18 and includes 3 different convolutional layers and 6 residual blocks, each of which contains 2 layers of ordinary spatial convolution and 2 transposed convolutions. First, the first part receives samples and labels as input, with a convolution kernel size of 7 × 7, a step size of 1, and a fill size of 3, and adds the instance normalization layer and Relu as the excitation function. The incentive function accelerates training and improves stability. The second and third layers are down-sampled with a convolution kernel size of 4 × 4, a step size of 1, and a fill size of 2 to obtain a 4 × 4 × 256 feature map. Next, the 3×3 is separated into 3×1 and 1×3, using spatially separated convolution in the middle part, whose aim is to reduce the number of parameters for network training. The final upsampling is performed using transposed convolution and the output uses the inverse function Tanh.

Discriminant Network

The function of the discriminator is to determine whether the image is stego or cover. During network training, the discriminator forces the generated encoded image to be as close as possible to the original image, using the same convolution kernel in the encoder as in decoder (Hu, et al., 2019). The architecture of the discriminator is shown in Figure 5. For the discriminator network, the entire network depth is 7. It inputs a true or false sample and determines whether it is true or false and the target domain it belongs to. The convolution kernel size is 4×4, the step size is 2, and the fill size is 1. The middle part is the implicit layer that enables the stable acquisition of symbolic features. The number of convolution kernels is 128, 256, 512, 1024, and 2048 in that order. And its output has two parts: the confrontation label and the classification label.

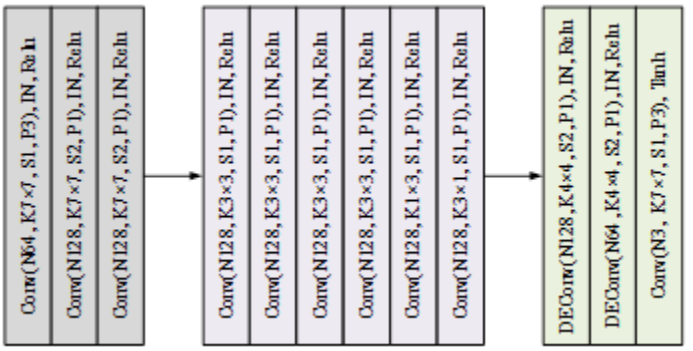
Figure 4. Generator structure



(a) 3D Incception V1

(b) I3D

Figure 5. Discriminator structure



Loss Function

In this paper, the loss function in the generative task of classical oil painting consists of the adversarial loss and L_2 (Ghodrati, et al., 2019) regularization constraint together, which is calculated as in Equation (4). Opposing loss uses Wasserstein distance as a measure of the distance between the generated data and the real data distribution to enhance the stability of the training.

$$L_{form} = L_{wgan} + \alpha L_2 \tag{4}$$

where L_{form}, L_{wgan}, and L_2 are the overall loss, adversarial loss, and L_2 regularization terms of the generative framework of the classical oil painting, respectively, and α is the regularization factor.

The original GAN uses purely a minimizing generator loss function during the training process to minimize the KL (Kapoor, et al., 2018) scatter between the generated distribution and the true distribution while maximizing the JS scatter of both, which leads to an unstable gradient, and for the distance between two samples that do not intersect at all, there is even a gradient disappearance. The defect of loss function leads to the difficulty of GAN training. When the discriminator performance is very good, the generator gradient disappears severely, and when the discriminator performance is insufficient, the generator gradient is not allowed to lead to unstable training. Wasserstein distance can calculate the expected lower bound in all possible joint distributions and calculate the optimal solution of its distance, even for completely disjoint data. Moreover, the gradient corresponding to WGAN varies almost linearly, solving the problem of gradient disappearance, and WGAN is calculated as shown in Equation (5).

$$W(P_r, P_g) = \inf_{\gamma \in \prod(P_r, P_g)} | E_{(x-y)\sim\gamma} | [|| x - y ||] \tag{5}$$

where $\prod(P_r, P_g)$ denotes the joint probability distribution of the real and generated data, i.e., $\sum_x \gamma(x,y) = P_r(x)$, $\sum_y \gamma(x,y) = P_g(y)$, $\gamma \in \prod(P_r, P_g)$.

EXPERIMENT

Experimental Environment and Evaluation Index

The experimental running environment is ubuntu 19.10 with 128G RAM, NVIDIA Tesla A100 GPU with 40G graphics memory. The experiment used Pytorch deep learning framework, development language Python 3.6.2, and Cuda environment NVIDIA CUDA 11.7 deep learning acceleration library. The network was trained with the stochastic gradient descent algorithm Adam in the experiments, with an initialized learning rate of 0.0001, and the model training loss curve is shown in Figure 6. In addition, to solve the model overfitting problem, Dropout is introduced to remove some neurons randomly, and Dropout takes the value of 0.3 in the paper. We can see from Figure 6, that when the number of model iterations Epoch is 20, the loss curves of both training and test sets tend to be smooth, and the loss values are below 0.2, indicating that the model has converged.

To prove the effectiveness of the algorithm in the paper, the model performance is evaluated using several mainstream evaluation metrics. The specific evaluation metrics include Accuracy, Precision, Recall, F1-score, and Time Overhead (TO) for action recognition of a single image. The calculated expressions are shown in Equations (6) to (9).

$$Accuracy = \frac{Tp + Tn}{Tp + Fp + Tn + Fn} \tag{6}$$

$$Precision = \frac{Tp}{Tp + Fp} \tag{7}$$

$$Recall = \frac{Tp}{Tp + Fn} \tag{8}$$

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \tag{9}$$

where *Tp* denotes positive cases predicted as positive cases; *Fn* denotes the number of counter-examples misreported as positive cases; *Fp* denotes the number of positive cases misreported as counter-examples, and *Tn* denotes the total number of detected counter-examples.

Analysis of Results

The confusion matrices successfully generated by the method in this paper for six different canvases in four sets of experiments are given in Figure 7. The rows of the matrix represent the labels of real oil paintings and the columns represent the optimization generated by the model in the paper. We can see that the accuracy rates for the generation of six different styles of oil paintings in the four sets of experiments were 72.51%, 75.01%, 76.45%, and 76.69%, respectively. Besides, the model in the paper can achieve a generation rate of 8ms/images. The above data can show us that the model in the paper performs stably on the results of multiple experiments, which further verifies that the

Figure 6. Loss value vs. accuracy curve: (a) comparison curves of loss function values, (b) accuracy curve

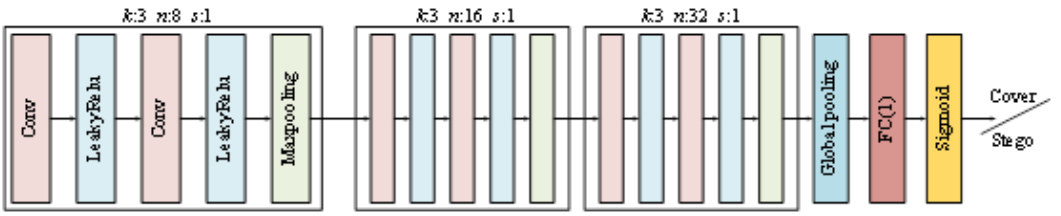
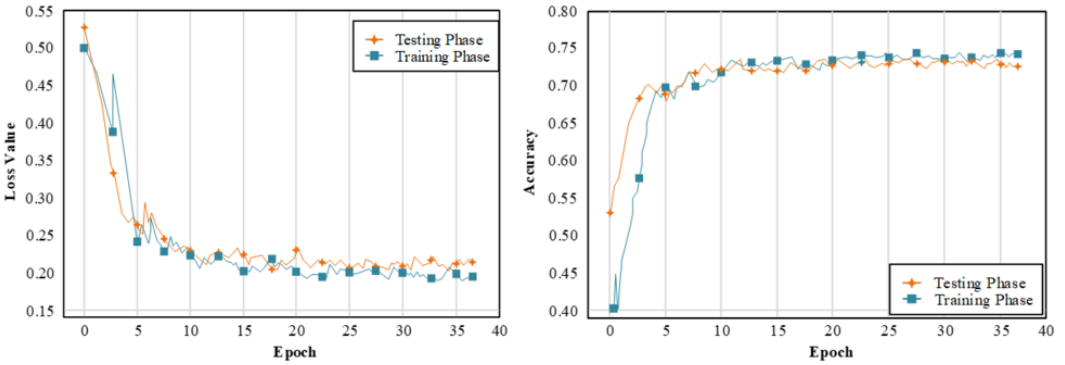


Figure 7. Confusion matrix



robustness and generalization performance of the model in the paper is good, and also has good real-time performance.

In addition, Figure 8 gives the comparison of the accuracy rate of oil painting generation for six compartments: landscape oil painting (A), flower oil painting (B), portrait oil painting (C), character custom oil painting (D), military oil painting (E), and history oil painting (F). In the paper, the average of the experimental results of 15 observers was selected as the final result.

As can be seen from Figure 8, the algorithm generates higher accuracy rates for all styles of optimization than the other comparative generation models, except for landscape oil painting and custom oil painting, which have a complex structure and strong personalization themselves, leading to low recognition rate. The generation accuracy rate of all other oil paintings reached over 73%. The above results show that the generated images conform to the viewing characteristics of the human eye for images. The main reason for this is that the model in the paper uses multimodal feature extraction techniques in multimedia feature processing, including still images, and video features, and encoding is performed in both temporal and spatial dimensions, which can effectively improve the encoding ability of features.

Comparison of Similar Related Works

The proposed model in the paper is compared with VAE (Zhang, et al., 2021), WGAN (Yang, et al., 2021), CGAN (Mishra & Herrmann, 2021), and InfoGAN (Cao, et al., 2022) under the same evaluation metrics and environment. The comparison is conducted using the same evaluation metrics and environment to ensure fairness and accuracy of the results. This comparison is crucial to assess the proposed model's performance and to determine its potential in various applications. Overall, the comparison results provide valuable insights into the effectiveness and limitations of the proposed model, as well as its potential for further improvements and applications. The experimental results are shown in Figure 9.

It can be seen that the combined advantages of this paper's model over the comparative models VAE, WGAN, CGAN, InfoGAN, etc. are obvious. Particularly, in the aspect of accuracy, the models in the paper improved by 2.01% (84.6 % vs. 86.3%) and 4.48% (82.6 % vs. 86.3%), respectively, compared to the best-performing VAE and InfoGAN. In terms of accuracy, this model improves by 1.30% (84.9 % vs. 86.1%) and 4.74% (82.2 % vs. 86.1%), respectively, compared to the best-performing VAE and InfoGAN models. In terms of recall, this model improves by 4.31% (83.5% vs. 87.1%) and 4.69% (83.2 % vs. 87.1%), and in terms of F1, the model in this paper improved by 4.32% (83.4 % vs. 87.0%) and 4.95% (82.9 % vs. 87.0%), respectively, compared to the best-performing VAE and InfoGAN models. The above experimental results verify that the model in the paper has good generative performance. The main reason for this is that the paper employs a multimedia feature extraction method that uses

Figure 8. Generation accuracy of different models for different canvases

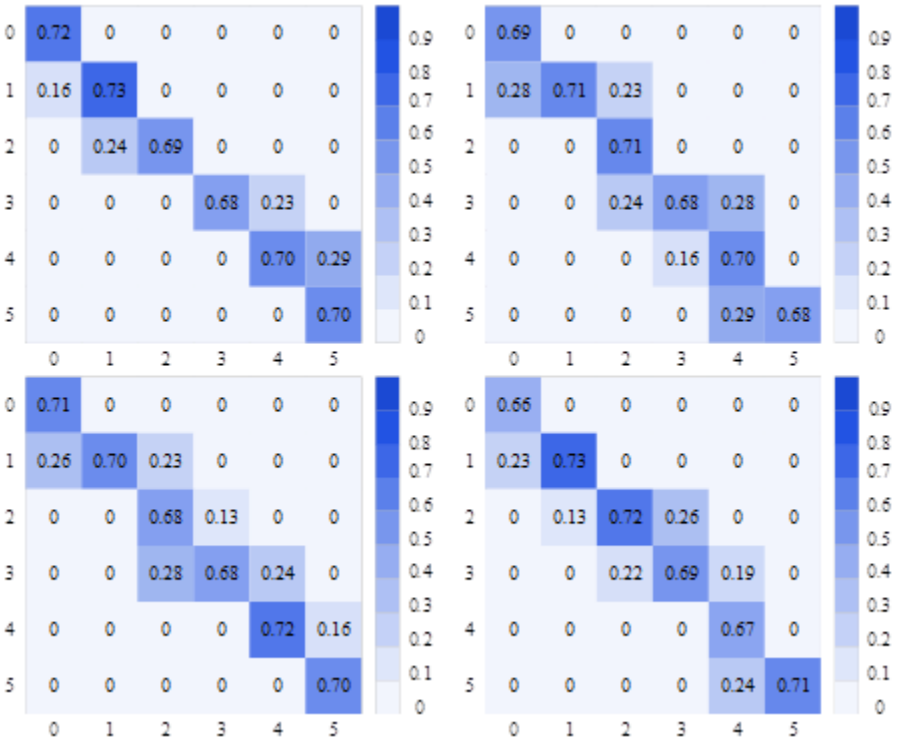
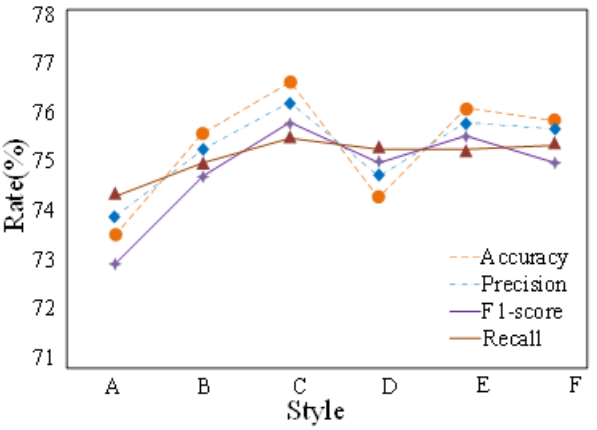


Figure 9. Comparison of the recommended performance of different models



pre-trained ResNet and channel attention to extract static features, dual-stream I3D features to extract dynamic features, and fused multimodal features to enhance the expressiveness of the features, and this effectively suppresses the edge information and focuses on the strong features that are better for the model classification performance.

In addition, to verify the time overhead of the classical oil painting generation model in this paper, tests are performed on the same data and environment. The results are shown in Figure 10.

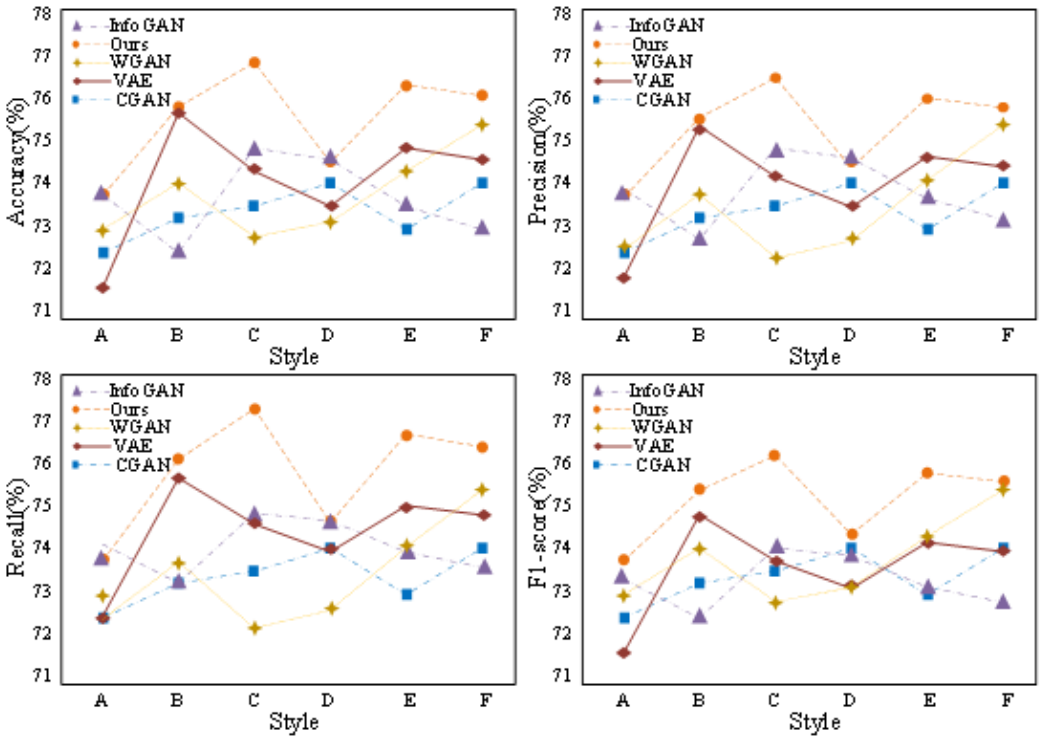
This shows that the model in the paper can achieve a generation rate of 6images/s, the WGAN model can achieve a generation rate of 6 images/s, and the CGAN model can achieve a generation rate of 5 images/s. The VAE model can achieve a generation rate of 9 images/s, and the InfoGAN model can achieve a generation rate of 7 images/s. The above data show that the model in this paper can also achieve a better recommendation success generation rate due to utilizing the pre-trained ResNet-50 as the backbone network. Additionally, feature extraction using both static and dynamic directions is significantly shorter in terms of time overhead compared to traditional single-direction feature extraction.

One of the variables that may have come into play in generating the results is the quality and quantity of the training data used for the generative model. While the authors made efforts to use a diverse range of classical oil painting styles, there may be some nuances in the styles that were not fully captured in the training data. Additionally, the generative model may be limited by the complexity of the input image and the style transfer algorithm used. Future research could explore alternative style transfer algorithms or incorporate additional data sources to improve the accuracy and realism of the generated images. Overall, while the results are promising, there is still room for improvement and further investigation.

CONCLUSION

The unique contribution of this paper lies in proposing a generative model of classical oil painting by combining multimedia data feature extraction techniques and generative adversarial networks in deep learning. The model achieves a new accuracy rate in the creation of oil paintings in six different styles, and it also has a better generation in real-time. However, there are limitations in the research, such

Figure 10. Comparison of the generation success rate of different models



as the interference of noise contained in the generated images, which affects the generation effect of landscape oil painting and custom oil painting with complex structures and strong personalized style. The research findings contribute to new knowledge by providing a novel and effective approach to generating classical oil painting pictures, which can be applied to teaching in universities and other fields. In future work, the generation effect of landscape oil painting and custom oil painting with complex structures and strong personalized style will be further improved by reducing the interference of noise contained in the generated images.

DATA AVAILABILITY

The figures used to support the findings of this study are included in the article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING STATEMENT

This work was supported by the 2022 Harbin University of Commerce Teacher ‘s “Innovation” Project and the 2022 Heilongjiang Province Philosophy and Social Science Research Project “Research on the Performance of Rock Color Painting in Harbin Red Architecture” (No. 22YSE441).

ACKNOWLEDGMENT

The authors would like to show sincere thanks to those techniques which have contributed to this research.

REFERENCES

Adnan, K., & Akbar, R. (2019). An analytical study of information extraction from unstructured and multidimensional big data. *Journal of Big Data*, *6*(1), 1–38. doi:10.1186/s40537-019-0254-8

Cai, Z., Xiong, Z., Xu, H., Wang, P., Li, W., & Pan, Y. (2021). Generative adversarial networks: A survey toward private and secure applications. *ACM Computing Surveys*, *54*(6), 1–38. doi:10.1145/3459992

Cao, D., Hou, Z., Liu, Q., & Fu, F. (2022). Reconstruction of three-dimension digital rock guided by prior information with a combination of InfoGAN and style-based GAN. *Journal of Petroleum Science Engineering*, *208*, 109590. doi:10.1016/j.petrol.2021.109590

Chen, T., & Yang, J. (2021). A novel multi-feature fusion method in merging information of heterogenous-view data for oil painting image feature extraction and recognition. *Frontiers in Neurorobotics*, *15*(709043), 709043. Advance online publication. doi:10.3389/fnbot.2021.709043 PMID:34322005

Gatys, L., Ecker, A., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414-2423). doi:10.1109/CVPR.2016.265

Ghodrati, V., Shao, J., Bydder, M., Zhou, Z., Yin, W., Nguyen, K. L., Yang, Y., & Hu, P. (2019). MR image reconstruction using deep learning: Evaluation of network structure and loss functions. *Quantitative Imaging in Medicine and Surgery*, *9*(9), 1516–1527. doi:10.21037/qims.2019.08.10 PMID:31667138

Gu, J., Sun, X., Zhang, Y., Fu, K., & Wang, L. (2019). Deep residual squeeze and excitation network for remote sensing image super-resolution. *Remote Sensing (Basel)*, *11*(15), 1817. doi:10.3390/rs11151817

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20–30.

Hossain, M. S., Bilbao, J., Tobón, D. P., Muhammad, G., & Saddik, A. E. (2022). Special issue deep learning for multimedia healthcare. *Multimedia Systems*, *28*(4), 1147–1150. doi:10.1007/s00530-022-00969-9 PMID:35844671

Hu, P., Peng, D., Sang, Y., & Xiang, Y. (2019). Multi-view linear discriminant analysis network. *IEEE Transactions on Image Processing*, *28*(11), 5352–5365. doi:10.1109/TIP.2019.2913511 PMID:31059440

Kahng, M., Thorat, N., Chau, D. H., Viégas, F. B., & Wattenberg, M. (2018). Gan lab: Understanding complex deep generative models using interactive visual experimentation. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 310–320. doi:10.1109/TVCG.2018.2864500 PMID:30130198

Kapoor, R., Gupta, R., Jha, S., & Kumar, R. (2018). Boosting performance of power quality event identification with KL Divergence measure and standard deviation. *Measurement*, *126*, 134–142. doi:10.1016/j.measurement.2018.05.053

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of GANs for improved quality stability and variation. In *Proceedings of International Conference on Learning Representations* (pp 1-26). Academic Press.

Ling, J., Zhang, K., Zhang, Y., Yang, D., & Chen, Z. (2018). A saliency prediction model on 360 degree images using color dictionary based sparse representation. *Signal Processing Image Communication*, *69*, 60–68. doi:10.1016/j.image.2018.03.007

Liu, B., Jiao, J., & Ye, Q. (2021). Harmonic feature activation for few-shot semantic segmentation. *IEEE Transactions on Image Processing*, *30*, 3142–3153. doi:10.1109/TIP.2021.3058512 PMID:33596173

Mishra, P., & Herrmann, I. (2021). GAN meets chemometrics: Segmenting spectral images with pixel2pixel image translation with conditional generative adversarial networks. *Chemometrics and Intelligent Laboratory Systems*, *215*, 104362. doi:10.1016/j.chemolab.2021.104362

Nandhini Abirami, R., Durai Raj Vincent, P. M., Srinivasan, K., Tariq, U., & Chang, C. Y. (2021). Deep CNN and deep GAN in computational visual perception-driven image analysis. *Complexity*, *2021*, 1–30. doi:10.1155/2021/5541134

Pambala, A. K., Dutta, T., & Biswas, S. (2021). SML: Semantic meta-learning for few-shot semantic segmentation☆. *Pattern Recognition Letters*, *147*, 93–99. doi:10.1016/j.patrec.2021.03.036

Qin, J., Luo, Y., Xiang, X., Tan, Y., & Huang, H. (2019). Coverless image steganography: A survey. *IEEE Access : Practical Innovations, Open Solutions*, *7*, 171372–171394. doi:10.1109/ACCESS.2019.2955452

Roy, A. G., Navab, N., & Wachinger, C. (2018). Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Transactions on Medical Imaging*, *38*(2), 540–549. doi:10.1109/TMI.2018.2867261 PMID:30716024

Rundo, L., Han, C., Nagano, Y., Zhang, J., Hataya, R., Militello, C., Tangherloni, A., Nobile, M. S., Ferretti, C., Besozzi, D., Gilardi, M. C., Vitabile, S., Mauri, G., Nakayama, H., & Cazzaniga, P. (2019). USE-Net: Incorporating Squeeze-and-Excitation blocks into U-Net for prostate zonal segmentation of multi-institutional MRI datasets. *Neurocomputing*, *365*, 31–43. doi:10.1016/j.neucom.2019.07.006

Sompairac, N., Nazarov, P. V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., Zhumadilov, Z., Barillot, E., Radvanyi, F., Gorban, A., Kairov, U., & Zinovyev, A. (2019). Independent component analysis for unraveling the complexity of cancer omics datasets. *International Journal of Molecular Sciences*, *20*(18), 4414. doi:10.3390/ijms20184414 PMID:31500324

Tang, X., Wang, Z., He, Q., Liu, J., & Ying, Z. (2020). Latent feature extraction for process data via multidimensional scaling. *Psychometrika*, *85*(2), 378–397. doi:10.1007/s11336-020-09708-3 PMID:32572672

Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., & Pinheiro, P. R. (2020). Covidgan: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection. *IEEE Access : Practical Innovations, Open Solutions*, *8*, 91916–91923. doi:10.1109/ACCESS.2020.2994762 PMID:34192100

Wang, G., Ye, J. C., & De Man, B. (2020). Deep learning for tomographic image reconstruction. *Nature Machine Intelligence*, *2*(12), 737–748. doi:10.1038/s42256-020-00273-z

Wang, H., Cao, J., Feng, J., Xie, Y., Yang, D., & Chen, B. (2021). Mixed 2D and 3D convolutional network with multi-scale context for lesion segmentation in breast DCE-MRI. *Biomedical Signal Processing and Control*, *68*, 102607. doi:10.1016/j.bspc.2021.102607

Wen, J., Fang, X., Cui, J., Fei, L., Yan, K., Chen, Y., & Xu, Y. (2018). Robust sparse linear discriminant analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(2), 390–403. doi:10.1109/TCSVT.2018.2799214

Yan, X., Cui, B., Xu, Y., Shi, P., & Wang, Z. (2019). A method of information protection for collaborative deep learning under GAN model attack. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *18*(3), 871–881. doi:10.1109/TCBB.2019.2940583 PMID:31514150

Yang, H., Lu, X., Wang, S. H., Lu, Z., Yao, J., Jiang, Y., & Qian, P. (2021). Synthesizing multi-contrast MR images via novel 3D conditional Variational auto-encoding GAN. *Mobile Networks and Applications*, *26*(1), 415–424. doi:10.1007/s11036-020-01678-1

Yang, H., Yuan, C., Li, B., Du, Y., Xing, J., Hu, W., & Maybank, S. J. (2019). Asymmetric 3d convolutional neural networks for action recognition. *Pattern Recognition*, *85*, 1–12. doi:10.1016/j.patcog.2018.07.028

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, *58*, 101552. doi:10.1016/j.media.2019.101552 PMID:31521965

Yu, Y., Huang, Z., Li, F., Zhang, H., & Le, X. (2020). Point Encoder GAN: A deep learning model for 3D point cloud inpainting. *Neurocomputing*, *384*, 192–199. doi:10.1016/j.neucom.2019.12.032

Zhang, Z., Cen, Y., Zhang, F., & Liang, X. (2021). Cumulus cloud modeling from images based on VAE-GAN. *Virtual Reality & Intelligent Hardware*, *3*(2), 171–181. doi:10.1016/j.vrih.2020.12.004

Zhu, J., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision* (pp 2223-2232). doi:10.1109/ICCV.2017.244

Zhuo Chen was born in Heilongjiang China, in 1989. From 2007 to 2011, she studied in Harbin University of Science and Technology and received her bachelor's degree in 2011. From 2011 to 2014, she studied in Harbin University of Science and Technology and received her Master's degree in 2014. Currently, she works in Harbin University of Commerce. She has published 10 papers, one of which has been published on Chinese core journal. Her research interests are included art layout, architectural landscape and national culture and economy.

Jianmiao Li was born in Heilongjiang, China, in 1980. From 1998 to 2002, he studied in Harbin Normal University and received her bachelor's degree in 2002. From 2013 to 2015, she studied in Harbin Normal University and received her Master's degree in 2015. Currently, he works in Harbin University of Commerce. He has published a total of 12 papers. There of which has been published on Chinese core journal. He research interests are plastic arts, decorative design, art education.