ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Visual Reconstruction of Cultural Symbols in Film and Television Animation Character Modeling Designs

Feng Long[1,*]

[1] Fuzhou University of International Studies and Trade, Fuzhou, Fujian, 350202, China

SUMMARY: *The aesthetic orientation of character modeling design is closely related to the final presentation effect and connotation expression of film and television animation. This paper takes the visual expression of cultural symbols in modeling design as the entry point, and chooses two forms of hue information extraction and gray processing to realize the digital compilation of cultural symbols. Combining YOLOv8m and image enhancement and other means to improve the performance of image detection in the symbol recognition link, using DBNet and CRNN to extract the text, accessing TSRAG for recognition, and adopting a weighted fusion decision-making mechanism to synthesize the two links, to establish an image-based dual-link logo recognition algorithm. On this basis, a compressed-aware image reconstruction method based on multi-scale inter-level feature fusion is proposed with compressed-aware and deep unfolding network as the core technology. The method optimizes the acquisition of local features and contextual information of cultural symbols by constructing a feature fusion proximal mapping module (MIFFPMM), and adds a new multi-scale inter-level feature fusion module in the deep reconstruction network to reduce information loss. The image reconstruction method can comfortably cope with the fluctuation of image sampling frequency, and the root-mean-square error is always lower than 0.25, and the lowest is only 0.01145, which is an effective visual reconstruction path for cultural symbols.*

KEYWORDS: *dual-link logo recognition; compressed-aware image reconstruction; character modeling design; cultural symbols*

1 Introduction

Film and animation both require culture and are one of the vehicles for cultural transmission [1]. Factors affecting the audience's perception of culture in animation in the cross-cultural environment reported in the literature [2] are, in order, clothing, character type, color, texture, and the audience's preference for character design is also accompanied by nationality. Therefore, character modeling design in animation is not only a form of artistic expression, but also a reflection of national aesthetic connotation. The introduction of cultural elements provides more possibilities for the styling design of animation characters, which makes the characters not only more visually appealing and diversified, but also more in-depth in cultural communication [3-5].

With the development of technology, the means of animation production is constantly enriched, from traditional two-dimensional animation to modern three-dimensional animation, the progress of technology allows designers to explore the expression of various materials more freely, and makes two-dimensional animation artificial three-dimensional sense enhancement,

*longfeng7522665@163.com
https://doi.org/10.65102/is2026038
Published: 30 April, 2026

to ensure that the quality of the character styling design improved [6-8]. In works such as “The Descent of Nezha the Magic Boy”, “Begonia the Big Fish”, and “White Snake: Origin”, Chinese film and television animation have integrated technology, culture, art, and ideas in depth, realizing the visual reconstruction of cultural symbols, and facilitating the dissemination of culture to the outside world [9-11]. However, many studies have indicated that due to regional differences, religious differences, differences in beliefs and other factors, the same element presents different cultural connotations, spiritual symbols and values in different countries, e.g., the dragon is a symbol of nobility and sanctity in China, while the dragon is a representative of evil in the West, which reflects the general dilemma of the translation of cultural symbols [12-15].

For the interpretation of cultural symbols in animation, scholars base on the theory of symbols. Literature [16] interpreted the art form symbols and cultural symbols of a number of film and television animations, among which The Return of the Great Sage and Kung Fu Panda both incorporate traditional Chinese characteristics to construct the cultural symbols of the characters. Literature [17] used Pierce's semiotic theory to analyze the animation “The Power of Mother” in the context of society, culture, and ideology, in which the audience used the character of mother as a symbolic representation and labeled it as a symbol of power. Literature [18] used Pierce's semiotic framework to analyze the visual meanings of the characters in the animated series of Nezha, and pointed out that visual symbols need to be combined with current cultural and psychological characteristics to create more resonant symbols. Reference [19] introduced formalist theory and visual semiotics into the "attraction" factor in local animation character design. Through the formation of "audience impression vocabulary" or "sensory vocabulary", it aims to understand and promote the formation of "attraction", thereby achieving a higher level of character design.

With the development of society and the iterative updating of technology, emerging technologies such as artificial intelligence have become important tools for the visual reconstruction of cultural symbols. Literature [20] reconstructed cultural symbols as dynamic graphics in a digital platform and generated a new short film animation based on this, and the survey showed that more than 60% of the audience supported this and believed that this animation promoted culture. Literature [21] decomposes and visually reconstructs Chinese cultural archetypes through generative adversarial networks and variational self-coders, and proposes a three-dimensional creation model, which solves the problem of homogenization of cultural symbols in animation by the process of decoding cultural archetypes, intelligent generation, and cultural verification. Literature [22] uses interactive technology to transform the cultural value of Helan Mountain petroglyphs into cultural symbols in animation, by analyzing the content associations and symbolism in the petroglyph motifs, and combining the basic knowledge system of the paintings, so as to create the animation.

This paper firstly points out the basic concept of image sampling and mapping, briefly describes the two methods of bitmap sampling and coding, and proposes the digital compilation technology of cultural symbols. Secondly, we sort out the design idea of dual-link logo recognition algorithm, focusing on the operation steps of the text detection and recognition module based on the logo a priori, and the mathematical operation of the decision fusion module. Based on the compressed perception and depth expansion network, the operation process of sampling network, initial reconstruction network and depth reconstruction network is explained in order to build the image reconstruction method of cultural symbols. At the same time, the classification and recognition performance comparison experiment is carried out to optimize the classification and recognition performance of the image reconstruction method. Finally, the effect and quality of the image reconstruction method are evaluated, and the application is analyzed.

2 Digital compilation of cultural symbols

2.1 Image Sampling Mapping

The most direct way for cultural symbols to be perceived is to create visualized images. The acquisition of raw images is usually divided into two cases: first, material cultural symbols can be obtained by photography or mapping to obtain image information, such as flowers, windows, landforms, etc.; second, non-material cultural symbols can be indirectly presented by textual markers, art works, etc., such as music scores, calligraphy and paintings. For this reason, after obtaining the original image, it is still necessary to carry out image sampling processing - simplifying or filtering multiple redundant information to form a coded source - is one of the most intuitive and efficient means of digital coding. This process usually consists of two steps from input to output, i.e., the original input image is first converted into a graphic file that is recognizable and expresses a specific meaning; and later these graphic files are transformed into a new form. Both of these transformations require certain technical aspects. In this process, the designer has to choose the appropriate coding scheme based on what he or she wants to convey; the size, color, and dimensions of the graphic are also taken into account. Designers need to use their subjective initiative, combined with the purpose of the design, through the design action to set up a set of morphological operation framework that reflects the design style. “Mapping” is a one-to-one correspondence between the coding source obtained through image sampling and the frame of morphological operation, which is also a more direct way of cultural symbol conversion.

2.2 Bitmap Sample Coding

The original image of a cultural symbol usually exists as a bitmap - a bitmap, also known as a dot image or raster image, consists of single dots called pixels (picture elements). These dots can be arranged and colored in different ways to form the image. Where the individual pixel units of a bitmap are direct reflections of cultural symbolic information - using this intuitive property, the valid information of the pixel units in a bitmap can be extracted for sampling and coding. Taking the RGB color model as an example, there are usually two methods of encoding:

First, extracting the hue information of each pixel, there will be (R,G,B) three dimensions of parameter variables. The information sampled by this method is more complex, with too many dimensions, which is no different from printing patterns if applied directly, and the coding logic is too straightforward, and is usually less used. The hue relationship can generally set the hue tendency in advance to serve as the evaluation weights, such as only picking up the red tendency of the region, to convey some kind of color emotion; or the hue information in the corresponding mapping range to extract the average value, such as mosaicization and so on.

Secondly, the picture is processed in grayscale, i.e., all the three variables of R, G and B are replaced by Gray variables, and the commonly used conversion formula is $\text{Gray} = R*0.299 + G*0.587 + B*0.114$. Due to the downgrading of the color information into one dimension, only the light and dark relationship of the picture is reflected, and no longer involves the multilevel value of the pixel, which makes the processing simple, and the processing and compression of the data is small, which can improve the efficiency of the translation afterwards, and it is very suitable for embodying the intention of the picture, so This is one of the most commonly used simplified operations.

In grayscale processing, if further dimensionality reduction of the information is required, further binarization can be used - according to a custom threshold, all pixels larger than the

threshold are set to 255 (white), and those smaller than the threshold are set to 0 (black). Local binarization also allows you to set one or more threshold intervals, within which the value becomes the specified color. This can namely reflect the obvious black-white-gray relationship. For example, it can be used to separate the contours of a graphic, extract the main simplified features, etc. .....

3 Image Reconstruction Methods for Cultural Symbols

3.1 Image-based dual-link identification algorithm

In the field of public sign recognition, relying only on a single image feature or text feature is often difficult to meet the demand for high-precision recognition in complex scenes. To solve this problem, this paper proposes an image-based dual-link logo recognition algorithm. The algorithm first accurately detects the logo symbol region; second, with the support of the symbol detection results, it uses the region of interest (ROI) to locate and focus on the text region, and completes the text recognition through DBNet combined with CRNN; then, based on the task-sensitive retrieval enhancement generating framework, it generates the final logo recognition results for the recognized text content; finally, based on the dual-link recognition results of the image and text, it generates the final logo recognition results with the detection and recognition confidence. Finally, based on the image and text dual-link recognition results, decision-level fusion is performed with the detection and recognition confidence as the weight to realize the accurate recognition of the logo. The whole algorithm flow is shown in Fig. 1, which includes the following three core modules: YOLO target detection and recognition based on sample expansion and progressive learning, text detection and recognition based on logo a priori, and decision fusion, and this section focuses on the text detection and recognition based on logo a priori and decision fusion module.

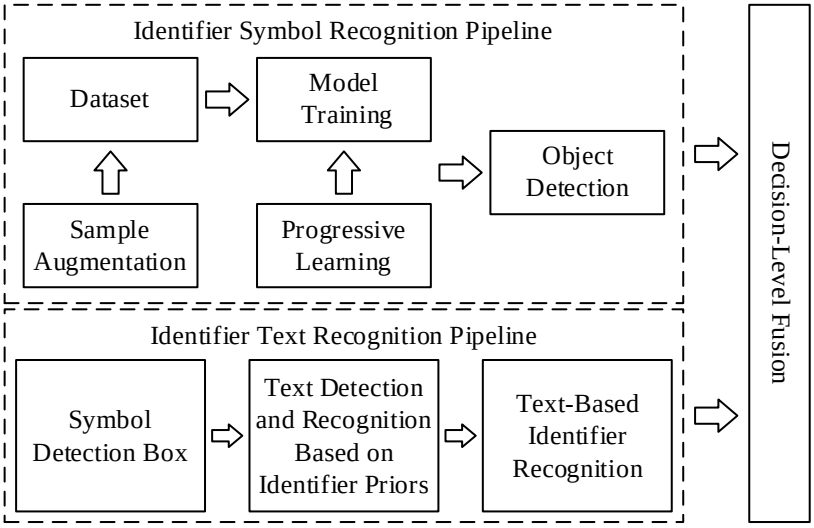


Figure 1: Flow chart of dual link identification algorithm based on image

3.1.1 Text Detection and Recognition Based on Marker A Priority

Aiming at the characteristics of strong spatial correlation between logo symbols and logo text in practical applications, this paper proposes a text detection and recognition module based on a priori knowledge of logos. This module can effectively correlate the symbol region in the

image with the corresponding text information, eliminate the interference of background cluttered text on logo recognition, and further improve the overall processing speed and accuracy. Different from the traditional full-image-by-image detection method, this scheme takes the detected symbol target region as the benchmark, and restricts the text detection and recognition operation to the local region (ROI). This design not only reduces the search space and improves the detection efficiency, but also effectively reduces the recognition error due to irrelevant text. Specifically, the module consists of four core steps:

(1) ROI region extraction

In the logo detection phase, the model outputs a series of target detection frames $B=\{B_1,B_2,...,B_n\}$, each detection frame B_i corresponds to a potential identifier number instance, containing the position parameter (x_i,y_i,w_i,h_i), which indicates the center point coordinates and width and height. Based on these detection frames, this paper extracts the local region of the image as the region of interest (ROI) for subsequent text detection.

Relevant standards clearly stipulate that the logo text should be arranged horizontally on the right side or below the identifier, and it is prohibited to place the logo text on the top of the identifier, and only a very few auxiliary texts of the security logo can be located on the left side of the identifier. Therefore, in this paper, the selection of ROI area is based on the left side, right side and bottom of the identifier. Specifically, for a detection frame expressed in the form of a top-left vertex and a bottom-right vertex, its target detection frame B_i can be expressed as equation (1):

$$(x_{i1},y_{i1},x_{i2},y_{i2}) \tag{1}$$

where equation (2):

$$x_{i2}=x_{i1}+w_i,y_{i2}=y_{i1}+h_i \tag{2}$$

In order to prevent the text from being truncated due to too tight cropping and to ensure the integrity of the detected text region, this paper expands a fixed number of times around the left side, right side and bottom of each detection box. The coordinate position of the upper left vertex of the target detection box remains unchanged in the process of expansion, and the position of the lower right vertex after expansion can be expressed as equation (3):

$$(x'_{i1},x'_{i2},y'_{i2})=(x_{i1}-M_l w_i,x_{i2}+M_w w_i,y_{i2}+M_h h_i) \tag{3}$$

where the left expansion multiplier $M_l=3$, the right expansion multiplier $M_w=5$, and the bottom expansion multiplier $M_h=1$. To ensure that the expanded ROI region does not exceed the boundary of the image, increase the boundary cropping operation. Set the image size as (W_{img},H_{img}). Crop the expanded coordinates as in equation (4)-(5):

$$x''_{i2}=\min(W_{img},x'_{i2}),y''_{i2}=\min(H_{img},y'_{i2}) \tag{4}$$

$$x''_{i1}=\max(0,x'_{i1}),y''_{i1}=\max(0,y'_{i1}) \tag{5}$$

The expanded ROI is denoted as B'_i as in equation (6):

$$B_i' = \left(x_{i1}'', y_{i1}'', x_{i2}'', y_{i2}'' \right) \tag{6}$$

(2) Localized text detection

Within each ROI, DBNet is used for text detection.DBNet has excellent fine-grained text boundary modeling capability, which can adapt to text instances of various shapes, and is especially suitable for dealing with complex scenarios such as curved text and italic text, which are common in logo images.

Different from the traditional regression text detector, DBNet predicts the pixel-level text probability map P with differentiable binarization threshold T and calculates the binarized output B as in equation (7):

$$B(x, y) = \sigma\left(k(P(x, y) - T(x, y))\right) \tag{7}$$

where σ denotes the sigmoid function, k is the magnification factor, and (x, y) is the pixel position.

The training process makes $B(x, y)$ approximate the actual text distribution, thus obtaining a high-quality text detection frame set $T = \{T_1, T_2, ..., T_n\}$, each T_j represents a detected text instance box. After the detection, a set of text candidate frames is obtained in each ROI.

It should be noted that when the local text detection stage fails to detect the effective text region, it can be determined that the logo image is missing text information, at this time the system will directly output the recognition results of the symbol.

(3) Text box screening strategy

In the ROI region corresponding to an identification symbol, there may be multiple text instances at the same time, so it is necessary to accurately match the detected text box with the symbol box. For this reason, this paper adopts the Hungarian matching algorithm to realize one-to-one optimal association.

Specifically, the cost function C between the symbol box B_i' and the text box T_j is defined firstly, where each element $C_{i,j}$ is defined as a composite distance metric as in Equation (8):

$$C_{i,j} = \lambda_1 \left(1 - IoU\left(B_i', T_j\right)\right) + \lambda_2 d_{center}\left(B_i', T_j\right) \tag{8}$$

where $IoU\left(B_i', T_j\right)$ denotes the intersection and concurrency ratio of the two, $d_{center}\left(B_i', T_j\right)$ is the Euclidean distance between the centroids of the two, and λ_1 and λ_2 are the weighting factors, set $\lambda_1 = 0.7$ and $\lambda_2 = 0.3$.

By minimizing the total cost matrix C, the Hungarian algorithm is used to complete the one-to-one optimal matching between the symbol box and the text box, so as to obtain a high-quality set of symbol-text correlation relations.

(4) Text Recognition

After completing the symbol-text matching, the CRNN model is used to recognize each matched text region.CRNN combines the capability of convolutional feature extraction and sequence modeling, which is able to deal with text inputs of variable lengths and complex morphology, and is suitable for the text recognition task in this scenario, which requires high accuracy.

Specifically, for each text detection frame T_j that is successfully associated with a symbol frame, a CRNN model is used for text recognition. The CRNN first extracts the feature representation $f(x)$ through a convolutional layer, and then performs sequence modeling through a bi-directional LSTM to ultimately output the probability distribution of the character $p(c|x)$ at each time step, and the overall process can be formalized as equation (9):

$$p(y|x)=\prod_{t=1}^{T} p(y_t|x) \tag{9}$$

where x is the input text image and y is the recognized character sequence.

Finally, based on the task-sensitive retrieval enhancement method, the final logo recognition result is generated for the recognized text content.

3.1.2 Decision integration

In the actual logo recognition task, there is often some uncertainty in the single image or text recognition results, especially in the case of complex environment (e.g., blurring, occlusion, low illumination) or logo variation (e.g., wear and tear, tilting, font change), the recognition accuracy is prone to decline. For this reason, this paper proposes a confidence-based decision-level fusion method, which makes full use of the complementary nature of the two information links, image recognition and text recognition, to effectively improve the accuracy and stability of logo recognition. The method takes the confidence scores output from the detection and recognition stages as the basis, considers the symbol detection and recognition confidence and text detection and recognition confidence respectively, and carries out the final logo determination by fusing the image recognition results and text recognition results at the decision level. At the same time, by analyzing the confidence distribution, difficult case mining can also be realized to provide support for subsequent model optimization and data enhancement.

Let the set of candidate categories output from the symbol detection and recognition stage be $C_{symbol}=\{(k_i,s_i)\}_{i=1}^{N}$, where k_i is the i th candidate category, s_i is the corresponding confidence score, and N is the number of candidate categories for the symbol branch. Similarly, the set of predicted results in the text recognition stage is $C_{text}=\{(k'_j,s'_j)\}_{j=1}^{M}$, where k'_j is the identification category obtained from parsing , s'_j is the confidence level of text recognition, and M is the number of candidate categories for text branching.

In order to synthesize the information of the two recognition paths, this paper adopts a weighted fusion decision-making mechanism. For each category k , the text synthesized confidence s_k is defined as equation (10):

$$s_k = w_{symbol}\cdot s_k^{symbol} + w_{text}\cdot s_k^{text} \tag{10}$$

where s_k^{symbol} is the confidence of category k on the symbol recognition path, set to 0 if no result; s_k^{text} is the confidence of category k on the text recognition path, set to 0 if no result; w_{symbol} and w_{text} are the fusion weights of symbol and text paths respectively, and w_{symbol} and w_{text} are set to 0.6 and 0.4 respectively in order to ensure the balanced contribution of the two paths in the experiment.

The category k^* corresponding to the final recognition result can be expressed as equation

(11):

$$k^* = \arg\max_k S_k \tag{11}$$

That is, the category with the largest combined confidence S_k is selected as the final logo recognition result.

3.2 Compression-aware image reconstruction based on multi-scale feature fusion

3.2.1 Sampling Networks and Initial Reconfiguration Networks

Same as the SAMNet method, the MIFFNet method also utilizes a convolutional layer for sampling and initial reconstruction, setting the network input as an image block of size $S \times S$, and employing a 1-layer unbiased convolutional layer to simulate the sampling process, and treating the convolutional kernel as a measurement matrix, the process of sliding and scanning the convolutional kernel over the picture is the process of compressed perceptual sampling. The random Gaussian matrix $A \in R^{M \times N}$ is reshaped into M learnable filters, each with kernel size $\sqrt{N} \times \sqrt{N} \times 1 = S \times S \times 1$, and then the sampling process is equation (12):

$$y^i = W_A * x^i \tag{12}$$

In Eq. (12), $x^i \in R^N$ denotes the ith image block, $y^i \in R^M$ denotes the measurements, $*$ is the convolution operation, and W_A is the filter weights.

The initial reconstruction reshapes $A \in R^{M \times N}$ into N filters, each of which has a size of $1 \times 1 \times M$, and then the tiles with a size of $N \times 1 \times 1$ are shuffled by pixel shuffling into $1 \times \sqrt{N} \times \sqrt{N}$, and finally get the initial reconstructed image, and the initial reconstruction process can be expressed as equation (13):

$$x_0^i = P\left(W_{A^T} * y^i\right) \tag{13}$$

In Eq. (13), x_0^i denotes the initial reconstructed image block and $P(\cdot)$ denotes the pixel shuffling operation.

3.2.2 Deep reconfiguration of the network

The deep reconstruction network consists of several repetitive stages, each containing two modules: the gradient descent module (GDM) and the attention proximal mapping module (SAPMM). Among them, multi-scale feature extraction and supervised attention are utilized in the SAPMM module to address the problem of inadequate feature extraction and lack of attention to global information. However, in the process of feature extraction and acquisition of global information, too many redundant features may affect the quality of the final feature representation. And irrelevant information such as noise or background interference can mask the truly valuable information, which results in the loss of important information leading to problems such as blurring of details, structural distortion and noise enhancement in the reconstructed image.

Therefore, the deep reconstruction network of MIFFNet method further improves the

proximal mapping module by unfolding the module through U-Net to introduce more amount of information, and at the same time, a multi-scale feature fusion module (MIFF) is designed in the proximal mapping module to solve the problem of too many redundant features and loss of important information.

The MIFFNet method uses the same gradient descent module (GDM) as the SAMNet method, and the newly proposed feature fusion proximal mapping module (MIFFPMM) for MIFFNet is described in detail. The module composition of the Feature Fusion Proximal Mapping Module (MIFFPMM) of MIFFNet is shown in Fig. 2, which is unfolded via U-Net. The U-Net has a U-shaped structure consisting of downsampling (encoder), upsampling (decoder) and jump connections. Expanding the proximal mapping module into a U-Net structure allows learning image information at different resolutions while retaining more high-frequency information through jump connections.

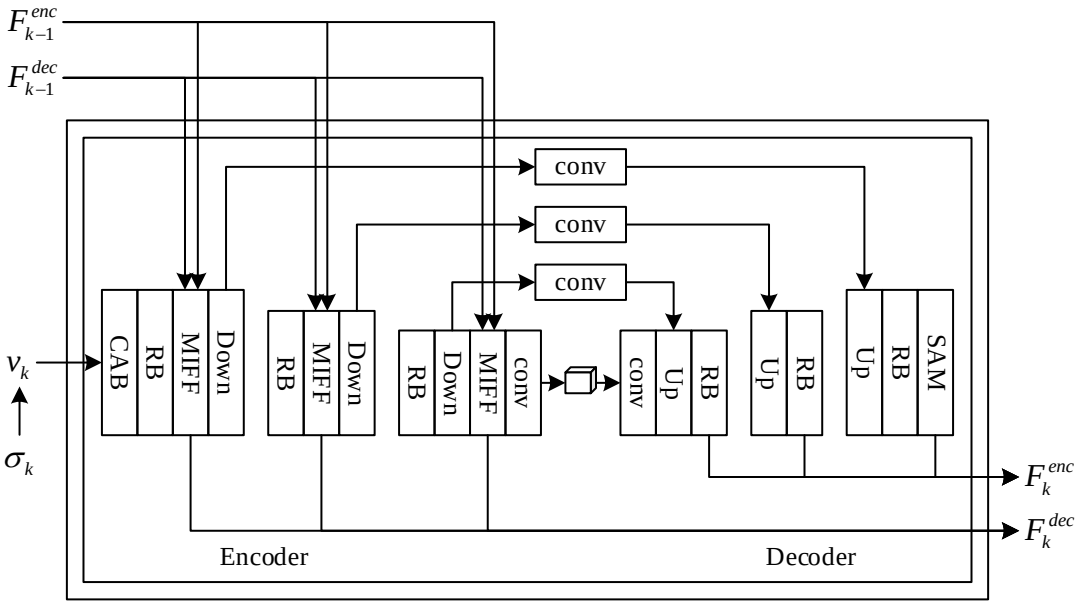


Figure 2: Feature Fusion Proximal Mapping Module (MIFFPMM)

The feature fusion proximal mapping module (MIFFPMM) first extracts shallow features with the help of channel attention block (CAB) to enhance the quality of the initial extracted features. After that, features at three different scales are further extracted with the help of residual block (RB) to obtain rich spatial structure information. In the downsampling process, 2×2 maximum pooling with a step size of 2 is used for feature dimensionality reduction, and bilinear upsampling is performed by convolutional layers, while global paths are introduced to motivate the network to avoid the low-frequency information so as to focus more on the detailed part of the image in order to solve the problem of insufficient feature extraction. At the end of the module, a supervised attention module (SAM) is utilized to further improve the focus on key feature regions and enhance detail recovery.

In addition the MIFFNet method is designed with a multi-scale inter-level feature fusion module (MIFF) at each scale of the encoder see Fig. 3.This module is designed to filter information-rich features and information-less redundant features by using thresholding to select the information weights to reduce feature redundancy to improve the feature extraction performance. Reducing feature redundancy helps to improve the efficiency of information transfer and retains valuable information, reduces the loss of important information, and effectively transfers it to the subsequent levels to enhance the feature capture and utilization

capabilities of the method.

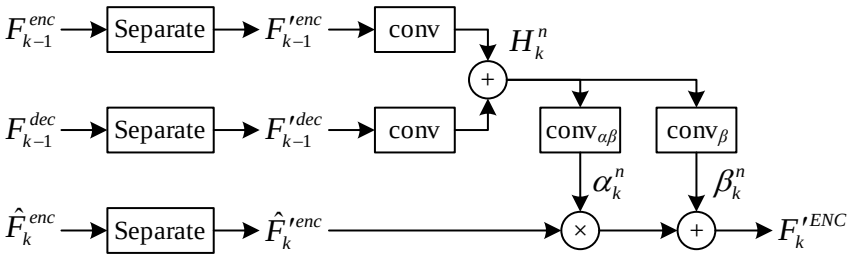


Figure 3: Multi-scale Inter-level Feature Fusion Module (MIFF)

Specifically, first, the features extracted from the encoder and decoder are denoted by $F_k^{enc} = \left\{ F_k^{enc \otimes n} \right\}_{n=1}^{3}$ and $F_k^{dec} = \left\{ F_k^{dec \otimes n} \right\}_{n=1}^{3}$, respectively, and the input features are normalized as in equation (14):

$$F_{out} = \alpha \frac{F - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta \tag{14}$$

In Eq. (14), μ and σ are the mean and standard deviation of the input features, ε is a small positive constant added for division stability, and α and β are trainable affine transformations.

Next, the information content of the different features is evaluated by measuring the spatial pixel variance of each batch and channel using the parameter γ of the trainable affine transformation in the normalization (GN) layer; the richer the information, the more pixel variations are reflected, and the greater the weight W_γ. This is used to indicate the importance of different features as in equation (15):

$$W_\gamma = \left\{ \omega_i \right\} = \frac{\gamma_i}{\sum_{j=1}^{C} \gamma_j} \tag{15}$$

Then, the weighted weights are mapped to the range of $(0,1)$ by the Sigmoid function and selected by the threshold, the weights above the threshold are the informative weights W_1, and the weights below the threshold are the non-informative weights W_2. There is equation (16):

$$W = Gate\left(Sigmoid\left(W_\gamma \left(F_{out} \right) \right) \right) \tag{16}$$

Finally, the input features are multiplied with W_1 respectively to obtain features that contain rich amount of information, while redundant features with little or no information are discarded as in equation (17):

$$\begin{cases} F_{k-1}^{\prime enc} = F_{k-1}^{enc} \odot W_1 \left(W > gate \right) \\ F_{k-1}^{\prime dec} = F_{k-1}^{dec} \odot W_1 \left(W > gate \right) \\ \hat{F}_{k}^{\prime enc} = \hat{F}_{k}^{enc} \odot W_1 \left(W > gate \right) \end{cases} \tag{17}$$

The above process is realized by the separation step in the Multi-scale Inter-level Feature Fusion (MIFF) module, whose network structure is shown in Fig. 4, where GN denotes the normalization operation, S denotes the Sigmoid activation function, and T denotes the Threshold threshold.

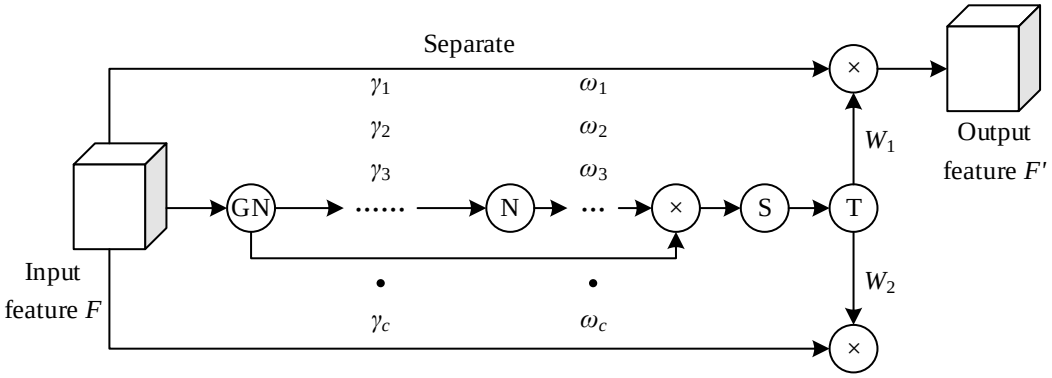


Figure 4: Separate Network

These features are then passed through separate 1×1 convolutional layers and merged by element-by-element addition, with the n th scale fusion result denoted as H_k^n at the k th stage. Next, two affine parameters $\left\{\alpha_k^n, \beta_k^n\right\}$ are utilized to convert the intermediate output $\hat{F}_k^{\prime enc}$ into the final information output $F_k^{\prime ENC}$. Mathematically, the inter-stage feature fusion process proposed in the paper can be defined as Eq. (18) feature representation:

$$\begin{cases} H_k^n = conv\left(F_{k-1}^{\prime enc} \right) + conv\left(F_{k-1}^{\prime dec} \right) \\ \alpha_k^n = conv_\alpha \left(H_k^n \right), \beta_k^n = conv_\beta \left(H_k^n \right) \\ F_k^{\prime ENC} = \hat{F}_k^{\prime enc} \odot \alpha_k^n + \beta_k^n \end{cases} \tag{18}$$

The above feature fusion process is the standard spatial adaptive normalization. Spatial adaptive normalization (SPADE) is a commonly used normalization method in image generation and image reconstruction tasks, which is mainly used to solve the problem of loss of important information caused by traditional normalization methods, such as batch normalization and instance normalization.

If F_k is utilized to represent the set of multi-scale encoder and decoder features, i.e., $F_k = \left\{ F_k^{enc}, F_k^{dec} \right\}$, the feature fusion proximal mapping module (MIFFPMM) of the MIFFNet method in the paper is denoted as Equation (19):

$$x_k, F_k = prox_{\theta_k} \left(v_k, F_{k-1} \right) \tag{19}$$

The θ_k in Eq. (19) denotes the parameter in the k th level.

4 Improvement and Evaluation of Cultural Symbolic Image Reconstruction

4.1 Improvement of Classification Performance of Reconstruction Methods

In order to improve the classification effect of the proposed reconstruction method on complex tasks, a combination of deep and shallow classification method is designed, which is combined with the method of this paper by introducing SVM with kernel function as histogram cross kernel.

From the data set of cultural symbols of film and television animation character modeling, the typical 10 categories of symbols (numbered 01-10 in order) are extracted as data set M. Data expansion is performed on data set M and negative samples are added to establish data set N. Then data set N contains the typical 5 categories of symbols (numbered 11-15 in order).

The dataset A+ is selected for experiments to compare with the methods with the best recognition performance based on shallow learning: (A1) SIFT+, (A2) HOG+ and (A3) RGB, as well as the combination of deep and shallow methods. The experimental results of multiple category recognition methods are shown in Table 1 for (C1) shallow learning, (C2) deep learning, and (C3) combination of deep and shallow. Overall (C1) shallow learning method has the weakest classification performance, both in terms of accuracy, recall and F1 metrics, which are within the (75.00,90.00)% interval. The (C2)deep learning represented by the proposed reconstruction method, on the other hand, improves, with the performance concentrated in the (85.00,95.00)% interval in the three metrics. While (C3) deep and shallow combination strategy, the classification performance is up to 100.00%, indicating that the combination of deep learning network and SVM using the method of classification and recognition performance is the best.

Table 1: The performance of various category recognition methods on the M dataset

Cate-gory	Accuracy(%)			Recall(%)			F1(%)		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
	A1-A3	Text-ual	Textual +SVM	A1-A3	Text-ual	Textual +SVM	A1-A3	Text-ual	Textual +SVM
1	84.09	93.83	98.75	78.99	92.67	95.14	88.57	92.21	96.98
2	82.99	90.06	100.00	83.4	87.15	97.84	84.46	92.83	95.63
3	87.67	92.01	99.87	78.3	87.3	100.00	89.65	88.23	97.52
4	76.95	90.34	100.00	80.26	85.55	97.57	80.75	91.7	95.13
5	84.36	87.24	98.05	80.03	86.42	9837	85.56	91.13	95.26
6	84.53	87.88	97.34	76.99	91.44	96.73	85.29	91.3	100.00
7	78.22	89.61	96.95	86.71	86.41	100.00	77.84	88.83	100.00
8	89.16	91.17	100.00	84.41	89.25	99.03	88.48	89.46	98.53
9	83.03	93.32	96.26	85.96	92.77	98.27	81.64	85.64	100.00
10	88.14	86.22	95.37	75.1	86.27	100.00	86.32	85.47	99.43

In order to further validate the recognition performance of the recognition system under the (C3) deep and shallow combination strategy, a more complex dataset N is chosen to do the

validation results are shown in Table 2.

Table 2: The performance of various category recognition methods on the N dataset

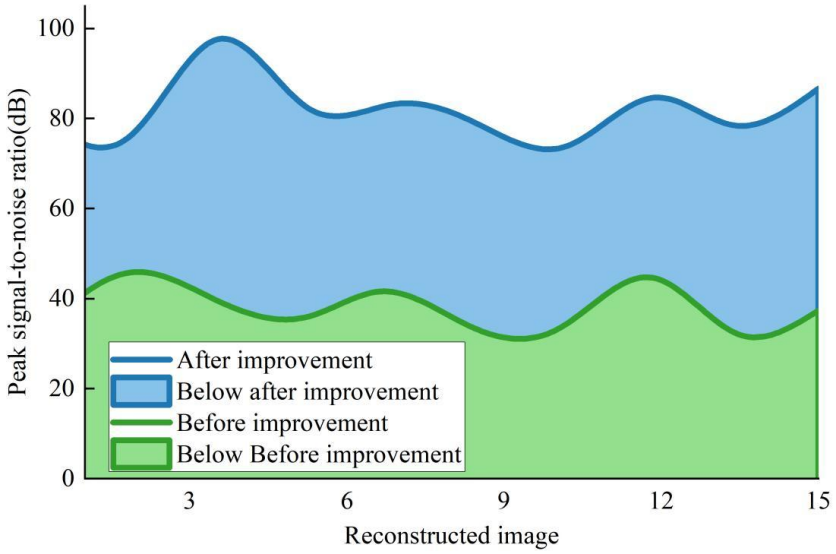
Cate-gory	Accuracy(%)		Recall(%)		F1(%)	
	C2	C3	C2	C3	C2	C3
	Textual	Textual+SVM	Textual	Textual+SVM	Textual	Textual+SVM
11	89.37	98.05	90.75	99.74	90.14	99.98
12	86.62	97.54	86.03	100.00	87.77	98.81
13	87.13	96.93	87.93	98.45	89.43	100.00
14	90.49	95.72	86.23	98.56	87.64	97.14
15	85.6	95.53	87.94	96.64	86.34	99.17

It can be seen that despite the increase in classification difficulty, the (C3) deep-shallow combination recognition method still achieves the highest performance of 100.00% in recall and F1 value optimization, and the accuracy ranges from 95.53% to 98.05%. The (C2) deep learning method is inferior to the (C3) deep-shallow combination strategy in all three indicators, with a maximum performance of only 90.75%. The reason is that SVM is extremely good at anti-noise and dimensional transformation, and can construct the optimal segmentation hyperplane in the feature space, so that the learning can reach the global optimum, so this paper adds the SVM method to the proposed reconstruction method.

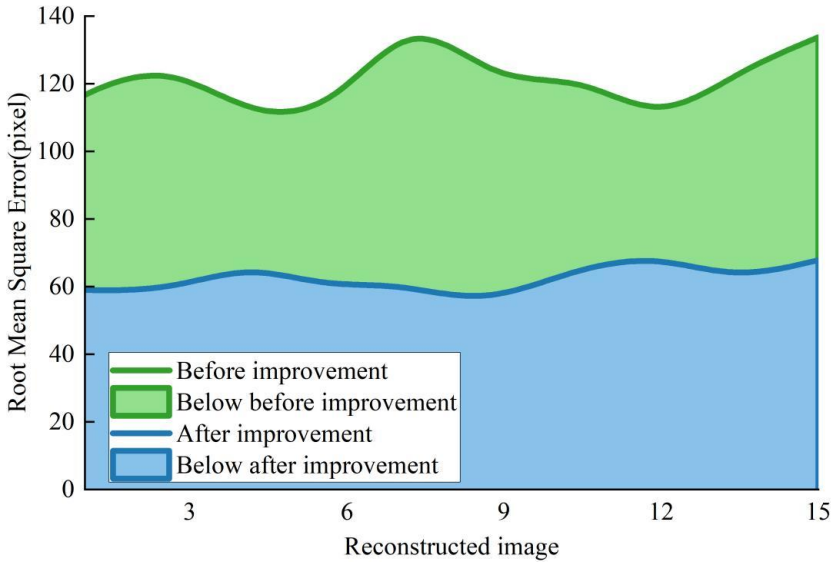
4.2 Effectiveness of the reconfiguration approach

4.2.1 Objective perspective analysis of reconfiguration effects

The stability and accuracy of the method proposed in this paper are verified by reconstructing the images of the target cultural symbols for 15 times. The 15 sets of peak SNR and rms error metrics are shown in Fig. 5(a)-(b). The peak SNR of the reconstructed images before improvement is always lower than 60dB, and the rms errors are higher than 110pixel and up to 133pixel; After the improvement of the method proposed in this paper, the peak signal-to-noise ratio of the reconstructed image is always stabilized at 75dB and above, which is a relative improvement of 15dB, and the mean square errors are all lower than 70pixel, which verifies that the method proposed in this paper can greatly improve the reconstruction quality of the cultural symbols.



(a) Peak signal to noise ratio

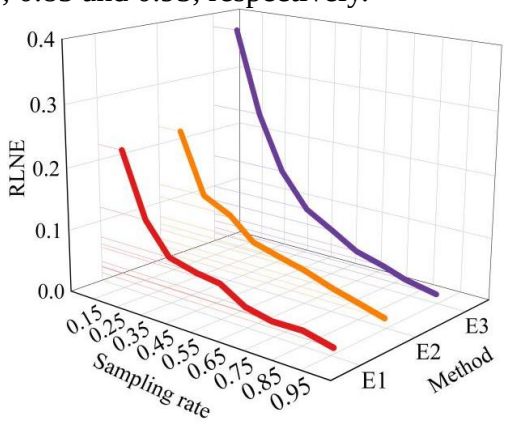


(b) Root-mean-square error

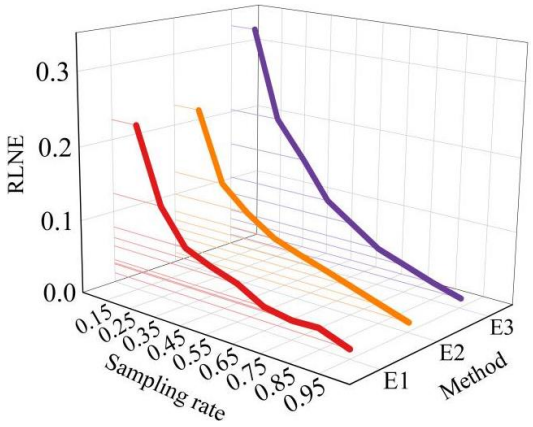
Figure 5: Objective index evaluation of image reconstruction effect

4.2.2 Effect of different data sampling frequencies on image reconstruction quality

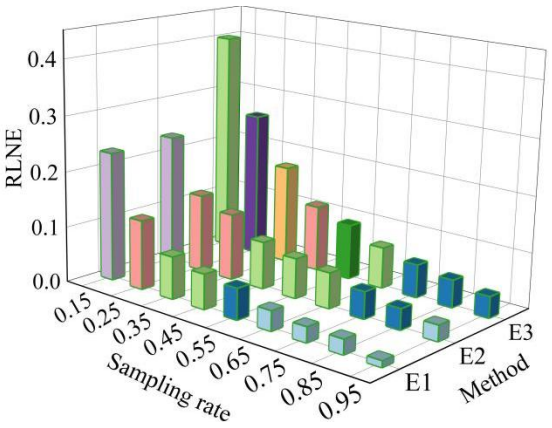
In order to explore the changes of the reconstruction quality of cultural symbol images in the reconstruction methods under different data sampling frequencies, this paper selects three Cartesian coordinate sampled images (numbered D1-D3 in order) and three 2-dimensional undersampled images (numbered D4-D6 in order) of a certain cultural symbol for the experimental objects. The relative l2 errors of combing with (E1) this paper reconstruction method for the six images are shown in Fig. 6(a)-(f) using (E2) GPBDCT and (E3) SIDCT methods as controls, and the sampling frequencies selected are 0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85 and 0.95, respectively.



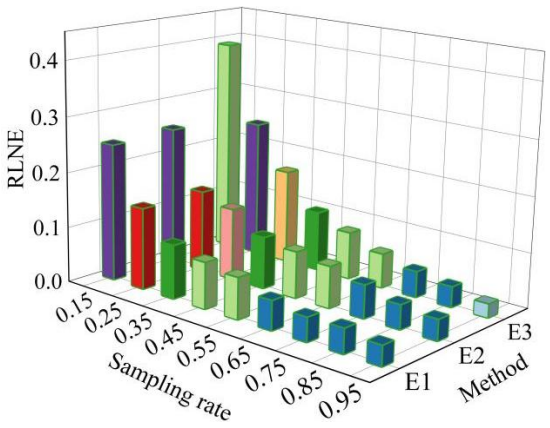
(a) D1



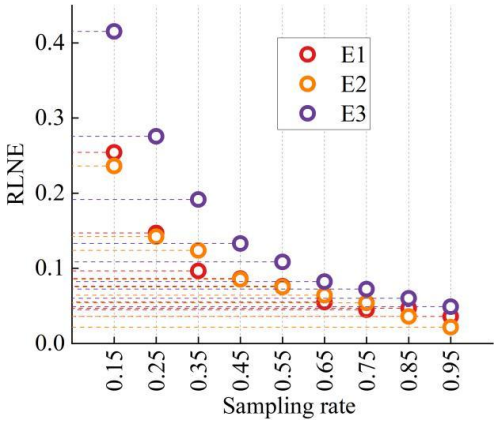
(b) D2



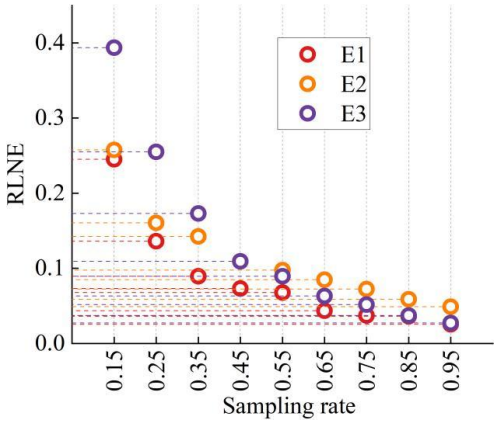
(c) D3



(d) D4



(e) D5



(f) D6

Figure 6: The variation of RLNE with sampling frequency under different methods

Comprehensively observing Fig. 6(a)-(f), the RLNE of all three methods decreases with the increase of sampling frequency, i.e., the higher the sampling frequency, the better the image reconstruction. Although the RLNEs of the three methods converge at 0.45 sampling frequency and beyond, the root mean square error of (E3) SIDCT method is always much higher than that of the other two methods at 0.15-0.45 sampling frequency, and the highest is up to 0.4153 in the starting point.(E2) GPBDCT method has a smoother performance of the RLNEs on six images, which is down to 0.01356.(E2) GPBDCT method has a smoother performance of the RLNEs on six images, and the lowest is up to 0.01356.(E3) SIDCT method has a lower RLNE than the other two methods. (E1) The overall RLNE of the reconstruction methods in this paper is <0.25, with a small fluctuation, always within the range of 0.2, and the lowest is only 0.01145<0.01356. Therefore, the method proposed in this paper is the best overall, with both strong stability and robustness.

4.3 Application of refactoring methods

The proposed reconstruction method is used to classify and recognize the cultural symbol dataset of film and animation character modeling, according to the number of occurrences in descending order is shown in Table 3, the top 10 are Chinese knot, panda, kung fu, calligraphy, taijiquan, twelve zodiac signs, Tang dynasty swords, xiao, bronzes, and lotus flower in the order of ranking. Among them, the Chinese knot symbol appeared the most 672 times, while the lotus

flower also appeared 326 times. The frequency of cultural symbols not only reflects the designer's preference and cultural heritage behind them, but also influences the theme tone of the film and animation works in a subtle way, and effective visual reconstruction can promote the quality of film and animation works.

Table 3: Statistics of high-frequency Chinese cultural symbols

Rank	Frequency	Chinese cultural symbols
1	672	Chinese knot
2	641	Panda
3	627	KungFu
4	558	Chinese Calligraphy
5	466	t'ai chi ch'uan
6	462	Twelve Chinese zodiac signs
7	401	Swordplay of the Tang Dynasty
8	359	Xiao
9	344	Bronze ware
10	326	Lotus

5 Conclusion

(1) This paper designs a dual-link logo recognition algorithm with YOLO target detection and recognition, text detection and recognition, and decision fusion as the core modules, combines compressed perception and deep unfolding network to establish an image reconstruction method applicable to the visual reconstruction of cultural symbols, and introduces SVM to enhance the classification and recognition performance.

(2) The image reconstruction method is able to achieve the optimal performance of recall and F1 value on data and up to 100.00%, while the accuracy is between 95.53% and 98.05%. The peak signal-to-noise ratio of the reconstructed images of the target cultural symbols is always at 75dB and above, and the mean square error is lower than 70pixel, and it shows superior robustness and stability compared with similar methods, and it can keep the RLNE <0.25, float 0.2, and the lowest can be as low as 0.01145 in the face of different sampling frequencies.

(3) The image reconstruction method is able to assist in the visual reconstruction of cultural symbols of film and animation character modeling design with high quality by means of powerful classification and recognition ability and excellent and stable reconstruction performance. Compared with similar methods that are widely used at present, its overall reconstruction effect is significantly superior, and it can provide reliable technical support for related fields.

About the Author

Feng Long was born in Fuzhou, Fujian Province, China in 1986. She obtained her bachelor's degree from Yunnan University and is currently pursuing a PhD at Universiti Teknologi MARA (UITM) in Malaysia. Her main research focuses on film and animation creation as well as digital media applications.

References

[1] Yusa, I. M. M., Ardhana, I. K., Putra, I. N. D., & Pujaastawa, I. B. G. (2023). Reality in animation: a cultural studies point of view. Eduvest-Journal of Universal Studies, 3(1), 96-109.

[2] Sattayasai, P., Jiang, P., & Tanaka, T. (2023). Cultural Parameters of Character Design–a Proposal of Cross-cultural Character Design Procedure for Visual Perception Achievement. International Journal of Asia Digital Art and Design, 27(1), 1-10.

[3] Wang, B. (2025). Archaeological Perspectives on the Application of Traditional Chinese Visual Elements as Cultural Symbols in Animation Creation. Mediterranean Archaeology and Archaeometry, 25(1).

[4] Sani, M. N. A., & Sin, N. S. M. (2024). Development of Character Costume Symbolism in Animation Folklore: A Systematic Review. Opportunities and Risks in AI for Business Development: Volume 1, 485-495.

[5] Feng, S. (2024). The Research on the influence of traditional Culture on the Role Shaping of Chinese Local Film and Television Animation. Cultura: International Journal of Philosophy of Culture and Axiology, 21(1).

[6] Arshad, M. R., Yoon, K. H., Manaf, A. A. A., & Ghazali, M. A. M. (2019). Physical rigging procedures based on character type and design in 3D animation. International Journal of Recent Technology and Engineering, 8(3), 4138-4147.

[7] Bouwer, W., & Human, F. (2017). The impact of the uncanny valley effect on the perception of animated three-dimensional humanlike characters. The Computer Games Journal, 6(3), 185-203.

[8] Wu, Y., & Chang, W. (2021). Research on the Character Creation of Chinese 3D Commercial Animation Films. Frontiers in Art Research, 3(6), 21.

[9] Xuefeng, S., & Sa-Ngiamviboon, A. (2024). The Cultural Representation of Nezha's Animation Image in the Context of Aesthetic Education. Journal of Roi Kaensarn Academi, 9(10).

[10] Sun, Y., & Hua, J. (2023). On strategies and effects of cross-cultural communication of Chinese mythological animated films—With Nezha and White Snake as examples. European Journal of Language and Culture Studies, 2(6), 6-14.

[11] Xiaoli, W. (2019). A Brief Analysis of the Application of Chinese Traditional Culture in Big Fish and Begonia. Journal of the Korea Entertainment Industry Association (JKEIA), 13(5), 67.

[12] Zhang, Q. S., Liu, Q. N., & District, B. (2020). A study of the differences between Chinese and Western cultures from the perspective of Hofstede's cultural dimension theory. East African Scholars J Edu Humanit Lit, 3(4), 125-128.

[13] Qi, J. C., & Ling, L. K. (2020). A Review on Western and Chinese Organization Culture: Similarities and Differences. Inti Journal, 2019.

[14] Chew, M. E., Ng, L. S., Jaafar, N. M., & Yeap, C. K. (2024). Understanding Oriental and Western Dragons in a Globalised World: A Cross-linguistic Study of Dragon-based Metaphorical Expressions in Chinese and English. 3L, Language, Linguistics, Literature, 30(4), 1-15.

[15] Maguth, B. M., & Wu, G. (2020). What Is the Difference Between the Chinese Dragon and Its Depiction in the West?. In Inquiry-Based Global Learning in the K–12 Social Studies Classroom (pp. 27-43). Routledge.

[16] Wang, P., & Chen, R. (2020). Symbolic Interpretation of Rare Animals in Movie and Anime Images. Revista Científica de la Facultad de Ciencias Veterinarias, 30(2), 908-918.

[17] Yunus, R. N., & Aswar, L. (2024). Semiotic Study Of The Animation Film Mother's Power: Representative Of Women's Power. International Journal of Multilingual Education and Applied Linguistics, 1(4), 67-79.

[18] WANG, W., & LI, W. (2024). CULTURAL PSYCHOLOGY AND ARTISTIC REPRESENTAMEN OF TRADITIONAL CHINESE IP ANIMATION CHARACTERS: UTILIZATION, INHERITANCE, AND INNOVATION OF HISTORY AND MEMORY. INNOVATION, 18, 49-59.

[19] Soikun, T. M., Ibrahim, A., & Asri, A. (2021). Finding “Appeal” Factors in Local Animation Character Design: Formalistic and Visual Semiotic Analysis (FVSA). Panggung, 31(2), 518110.

[20] Urgancı, E. M. (2025). The Use of Cultural Symbols in Digital Media: An Evaluation Through Motion Graphics. Selçuk Üniversitesi Sosyal Bilimler Meslek Yüksekokulu Dergisi, 28(1), 245-271.

[21] Xing, B. (2025). Deep learning driven recreation of traditional ethnic elements in animation works from the perspective of prototype theory. International Journal of Information and Communication Technology, 26(16), 38-52.

[22] Hongjie, H., Bunlikhitsiri, B., & Panthupakorn, P. (2024). The Symbol Transformation of Helan Mountain Rock Paintings: Animation Creation Based on Interactive Technology. Journal of Roi Kaensarn Academi, 9(8), 1407-1422.