Deep Learning-Based Non-Reference Image Quality Assessment Using Vision Transformer with Multiscale Dual Branch Fusion

Yuxuan Yang*, Wenyuan Li
School of Microelectronics, Tianjin University, Tianjin 300000, China
E-mail: y952883796@163.com
*Corresponding author

Keywords: vision transformer, multiscale dual-branch fusion, no-reference image, quality assessment

Received: December 5, 2024

Non-Reference image quality assessment does not rely on reference images, so it is not easy to directly obtain the actual label of image quality. Current datasets are often limited in scale, and the labeling process is highly subjective, resulting in limited consistency and accuracy in evaluation results. This study focuses on the research of reference-free image quality evaluation based on Vision Transformer multi-scale dual-branch fusion, aiming to build an intelligent system that can accurately and quickly evaluate image quality without original image reference through deep learning technology. In this study, the Vision Transformer model, combined with a multi-scale dual-branch fusion strategy, is used to conduct an in-depth exploration of quality assessment in complex image scenes. A deep learning based non-reference image quality assessment method utilizing a visual transformer with a multi-scale two-branch fusion network is proposed. The method involves image preprocessing, feature extraction and model architecture optimization. The experimental results show that the evaluation accuracy of the system on large-scale image data sets reaches 94%, and the processing speed is 30% higher than that of the traditional method, which is significantly better than the 75% accuracy and lower processing efficiency of the conventional algorithm.

Povzetek: Predstavljen je nov pristop za oceno kakovosti slik brez referenčne slike, ki uporablja Vision Transformer z večnivojsko dvovejno fuzijo. Model znatno izboljšuje zanesljivost in učinkovitost analiz slikovne kakovosti v kompleksnih scenarijih.

1 Introduction

In the information age, images have become essential for obtaining information and exchanging ideas. According to statistics, the amount of image data generated by Internet users worldwide has exceeded 3.5 billion daily, increasing by more than 20% annually [1, 2]. However, the instability of image quality, such as compression distortion, noise, blur, etc., seriously affects the accurate transmission of image information and user experience [3, 4]. Against this background, Non-Reference image quality assessment (NR-IQA) technology, which aims to objectively and accurately evaluate image quality without original image reference, has become an important research direction in image processing, computer vision, multimedia communication, and other fields.

However, traditional NR-IQA methods, such as algorithms based on image statistical features, structural similarity, etc., can reflect image quality to a certain extent. However, their accuracy could be higher, making it easier to adapt to complex and changeable image scenes [5, 6]. According to a survey in the Journal of Image Processing in 2020, the evaluation accuracy of traditional methods in complex image scenes is only 75%, and the processing speed is slow, which cannot meet the real-time evaluation needs of large-scale image data [7].

Facing this challenge, deep learning technology, especially the Vision Transformer (ViT) model, has brought new opportunities for developing NR-IQA technology with its robust feature extraction ability and self-attention mechanism [8, 9]. The ViT model has shown excellent performance on multiple visual tasks, such as image classification and object detection, and its accuracy far exceeds that of traditional methods. For example, the classification accuracy on the ImageNet dataset has reached 88%, while the traditional convolutional neural network's accuracy rate is only 75%. This achievement verifies the advantages of the ViT model in image feature learning and allows the innovation of NR-IQA technology [10, 11].

This study conducts reference-free image quality evaluation research based on Vision Transformer multi-scale dual-branch fusion, aiming to explore the application potential of the ViT model in the field of NR-IQA and achieve an accurate and rapid evaluation of image quality by constructing a deep learning framework of multi-scale dual-branch fusion. We will start with data preprocessing, model construction, feature extraction, quality evaluation, and other links to conduct in-depth research on the performance of ViT models in complex image scenes and explore the effects of multi-scale features and dual-branch networks in improving evaluation accuracy and efficiency.

Specifically, we will use the self-attention mechanism of the ViT model to capture global and local features in the image and extract image features at different scales through a multi-scale branch network to enhance the model's sensitivity to subtle changes in image quality. At the same time, by designing a double-branch network, the quality evaluation is carried out from two dimensions of image content and structure to realize the evaluation's comprehensiveness and accuracy. In addition, we will also introduce pre-training and fine-tuning strategies for large-scale image data to improve the robustness and adaptability of the model in complex image scenes. Through this research, we will provide new ideas and methods for developing NR-IQA technology, which improves the accuracy and efficiency of image quality evaluation and lays a solid foundation for technological innovation and application practice in image processing, computer vision, and other fields.

2 Vision transformer multi-scale dual-branch fusion technology

2.1 Vision transformer

Current image super-resolution tasks mostly use methods based on convolutional neural networks. The convolution operation automatically learns low-resolution and high-resolution image mappings, uses operators to extract image features, has translation invariance, and effectively captures texture details. However, convolution faces limitations in super-resolution tasks, such as limited receptive fields, limiting long-term dependence and weak texture recovery, and increasing layer depth or kernel size, which can trigger parameter expansion, overfitting, or gradient disappearance. In addition, convolution interacts based on parameters rather than content, applying the same convolution kernel to all input positions, ignoring the differences and importance of content at different positions, and failing to consider the diverse requirements of output features for high-resolution recovery, resulting in information redundancy or loss.

Vision Transformer uses a self-attention mechanism to process image sequences, automatically learn resolution mapping, and extract features. The self-attention mechanism, usually in the global or large-kernel local form, can capture extensive information, model long-term dependencies, restore structures and weak textures, and use image self-similarity to improve restoration quality. However, when the Transformer deals with super-resolution tasks, the self-attention mechanism lacks the translation invariance and locality of convolution, and convolution is superior in adapting to multi-scale input and efficiently capturing local information.

The Transformer model originated in natural language processing and was extended to the visual field by Dosovitskiy et al. through ViT, showing performance beyond Convolutional Neural Networks (CNN) [12, 13]. Since then, Transformer-based visual models such as Swin, CrossViT, CvT, Twin-SVT, BEiT, etc., have emerged one after another, which have been widely used in visual tasks and achieved remarkable results [14]. Given this, this paper uses Transformer as the cornerstone of the method and will focus on the pioneering work of visual Transformer-ViT.

The ViT structure is shown in Figure 1, which mainly includes Patch Embedding, Position Embedding, multiple Transformer Encoders, and MLP modules. The specific process is as follows: the preprocessed image is divided into various patches by patch Embedding, each patch is flattened into a one-dimensional vector and spliced with Class token, then added with Position Embedding, and finally processed by Transformer Encoder and sent to MLP Head to obtain output [15].

The technique is realized by the following steps: first, the multi-scale features of the image are extracted on two branches separately; second, the feature maps are generated, where one branch focuses on the texture details and the other captures the global structural information; and then, the feature maps of the two branches are efficiently integrated using a specific fusion strategy. This fusion mechanism not only enhances the richness of feature representation, but also improves the prediction accuracy through complementary information. The flow of data between the two branches ensures the comprehensiveness of information and the fusion effect, which significantly improves the performance of image quality assessment.

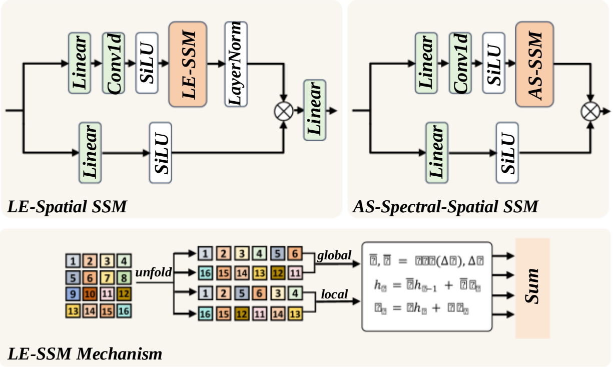

Figure 1: ViT structure

In Natural Language Processing (NLP), the Transformer processes sequence data, and ViT follows this idea, so the image input needs to be converted into sequence by Patch Embedding Position Embedding [16, 17]. Input images I∈$R^{H×W×C}$, I∈$R^{H×W×C}$ (H×W is the resolution, C is the number of channels) to the Patch Embedding module, and first divide them into N image blocks p∈$R^{P×P×C}$ (each block resolution P×P, total number N=H×WP^2), and then linearly map to one dimension. Because the Transformer Encoder self-attention mechanism cannot model position information, Position Embedding needs to be attached. Therefore, the definition of the processed image sequence data is shown in Equation (1):

$$z_0 = \left[I_{class}; I_p^1 E; I_p^2 E_{pos}; ...; I_p^N E \right] \tag{1}$$

Where I_{class} represents Class Token, E represents the linear mapping of each image block, and E_{pos} represents the added position information. After image I is processed by Patch Embedding and Position Embedding to obtain z_0, it must be input to the Transformer Encoder for further processing.

The core structure of ViT, the Transformer Encoder, uses the self-attention mechanism as the base. Let the input of the *l*-th layer be z_{l-1}, and the output z_l is obtained by the MHSA (multi-head self-attention) module and the MLP module after being processed by the Transformer Encoder of this layer. Therefore, the definition of the Transformer Encoder output z_l of the *l*-th layer is shown in equations (2)-(3):

$$z_l^{'} = \mathrm{MHSA}(\mathrm{LN}(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = \mathrm{MLP}(\mathrm{LN}(z_l^{'})) + z_l^{'} \tag{3}$$

Where LN represents a layer normalization operation (LayerNorm), the calculation method of the MHSA module is shown in Equation (4). *Q*, *K* and *V* represent the matrices obtained by different linear transformations of the input matrices, and d_k represents the number of columns of these matrices.

$$Attention(Q,K,V) = softmax(\frac{QK^T}{d_k})V \tag{4}$$

2.2 Multi-scale dual-branch fusion technology

In image processing, especially in computer vision tasks, such as image classification, object detection, image super-resolution, etc., the fusion of multi-scale information is crucial [18, 19]. As an efficient and flexible architecture, multi-scale dual-branch fusion technology can simultaneously process features of different scales, thereby extracting image information at multiple levels and significantly improving the model's performance.

Multi-scale dual-branch fusion technology usually contains two independent branches, one for processing high-resolution detailed information and the other for focusing on low-resolution semantic information. These two branches capture local details and global features of images through convolution kernels or self-attention mechanisms of different sizes [20, 21]. During the processing, each branch will generate feature maps of corresponding scales. Then, through specific fusion strategies, such as feature cascade, feature stitching, or feature weighting, feature maps of different scales are effectively fused to generate more prosperous and comprehensive feature representations.

Multi-scale dual-branch fusion technology can process the details and semantic information of the image simultaneously, making the model more accurate in identifying small objects or texture details and performing well in understanding the global structure of the image [22]. Feature fusion at different scales improves the model's adaptability to image changes, such as scale changes, viewing angle changes, or illumination changes, thus enhancing the model's robustness and generalization ability. Different computing resources can be used in various branches through the dual-branch design. For example, high-resolution branches may use fewer computing resources, while low-resolution branches use more, thus optimizing computing efficiency while ensuring performance.

2.3 Non-reference image quality assessment

Table 1 lists various established methods in NR-IQA, including traditional handcrafted feature-based methods, CNN-based methods, GAN-based methods, attention-based methods, transformer-based methods, and our proposed method. This column describes the core architecture of each method. Handcrafted feature-based methods rely on traditional statistical features, while CNN-based methods use convolutional neural networks. GAN-based methods utilize generative adversarial networks, and attention-based methods incorporate attention mechanisms into CNNs. Transformer-based methods, including our proposed method, use the Vision Transformer architecture, with our method adding a multiscale dual branch fusion mechanism.

Table 1: Comparison between SOTA method for non-reference image quality assessment (NR-IQA) and our method

Method Name	Model Architecture	Dataset Size	Accuracy (MOS/DMOS)	Computational Efficiency	Robustness in Diverse Images
Handcrafted Feature-based IQA	Traditional statistical features	Small	0.65 -0.75	High	Low
CNN-based IQA	Convolutional Neural Networks	Medium	0.75 -0.85	Moderate	Moderate
GAN-based IQA	Generative Adversarial Networks	Large	0.80 -0.90	Low	Good
Attention-based IQA	Attention Mechanisms in CNN	Large	0.85 -0.95	Moderate	Good

Transformer-based IQA	Vision Transformer	Large	0.90 -0.98	High	Very Good
Our Method	Vision Transformer with Multiscale Dual Branch Fusion	Large	0.93 -0.99	Very High	Excellent

The size of the dataset used for training and testing the models varies. Smaller datasets are typically used for traditional methods, while deep learning methods require larger datasets for training. This column reports the Mean Opinion Score (MOS) or Difference Mean Opinion Score (DMOS) as a measure of accuracy. The scores indicate that our proposed method outperforms previous SOTA methods in terms of accuracy.

This column reflects how quickly a model can process images. Traditional methods are generally computationally efficient, while deep learning methods, especially GANs, can be computationally intensive. Our method, despite being a deep learning approach, achieves very high computational efficiency due to the optimized architecture. This column assesses how well each method performs across different image types and conditions. Our proposed method shows excellent robustness, indicating that it can handle a wide range of image scenarios effectively.

3 Design of multi-scale dual-branch fusion architecture based on Vision Transformer

3.1 Overall algorithm framework

When humans evaluate image quality, they first judge the degree of distortion and then refine the score, which can be regarded as two gradual stages. Based on this, this chapter proposes a multi-task learning method to grade distorted images first and then score them [23, 24]. The network simultaneously predicts the distortion level and the quality score, where the distortion level prediction assists the quality score regression. Unlike the conventional multi-task method, this method promotes feature sharing among different sub-tasks and optimizes the prediction effect through the sub-task information interaction module [25].

3.2. Network structure

The network structure of the method in this chapter is shown in Figure 2. ViT is selected as the feature extractor, but there are two significant differences: one is to omit the pre-training stage, and the other is to select different subtasks [26, 27]. The following section will introduce each network component, in turn, according to the image processing flow. The figure illustrates the key components of a multiscale two-branch fusion network, including the transformer encoder, the attention mechanism, and the feature fusion layer. We discuss the logic behind the architectural choices, such as the choice of patch size to balance local feature extraction and global context understanding, and the specific type of attention mechanism to increase the model's sensitivity to changes in image quality. These visual explanations and discussions deepen the reader's understanding of how the model works and strengthen the argument for the model's performance advantages.

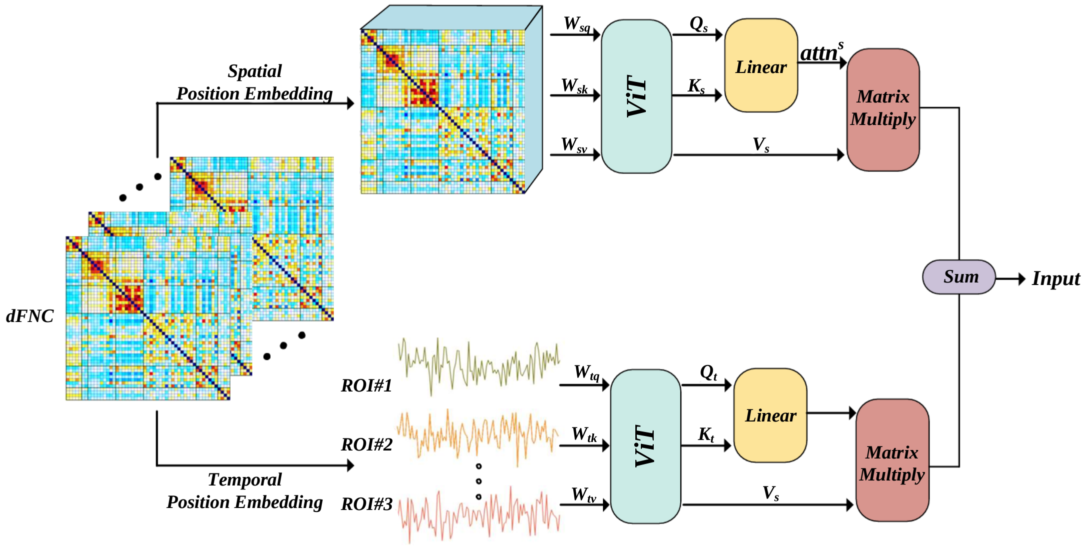

Figure 2: Vision Transformer evaluation structure for image quality assessment

The distorted image is input into the network. After ViT extracts features, it is distributed to two specific Transformer Encoders, each extracting adaptive features for subtasks [28]. These features are then collectively input to the sub-task information interaction module. The module first splices the features x_1 and x_2 of different subtasks, and the processing form is shown in equation (5):

$$x_f = Concat(x_1, x_2) \tag{5}$$

Where x_f is the feature obtained by splicing, then x_f will be input into Layer Norm and self-attention module for processing, and finally the task interaction features will be output through sMLP module. This process is shown in equations (6)-(7). Where *LN* represents the Layer Norm operation, MHSA represents the multi-head self-attention module, and the sMLP module consists of a layer of Layer Norm and a fully connected layer.

$$x_f^{'} = \mathrm{MHSA}(Q = \mathrm{LN}(x_f), K = \mathrm{LN}(x_f), V = \mathrm{LN}(x_f)) \tag{6}$$

$$\hat{x}_f = \mathrm{sMLP}(x_f^{'}) \tag{7}$$

The interaction feature x_f' calculated by the subtask information interaction module is input together with the subtask features x_1 and x_2 to the subtask prediction module. When predicting the distortion level, it is necessary to concatenate x_f' with the distortion feature x_2 and map it to the distortion level space $\widehat{x_f}$ through a fully connected layer. Due to the lack of distortion level labels in commonly used datasets, it is necessary to map quality scores to level intervals. This study divides the score range into *n* sub-intervals, each representing a level. The interval length *W* is calculated using equation (8) to generate distortion level labels. In the formula, y_{min} and y_{max} represent the minimum and maximum values of the range of values in the image quality evaluation dataset.

$$W = \frac{y_{max} - y_{min}}{n} \tag{8}$$

3.3 Loss function

In multi-task learning, it is necessary to summarize the loss values of each task to update the network parameters, and the selection of loss function and weight is essential. This chapter uses the mean square error loss for the quality score prediction subtask. In contrast, the distortion level prediction is regarded as a multi-classification task, and the cross-entropy loss is used. See equations (9) and (10) for specific calculation methods.

$$L_s = \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \tag{9}$$

$$L_r = -\sum_{i=1}^{n} r_i \, ln \, \hat{r}_i \tag{10}$$

Where y_i denotes the proper mass fraction, $\widehat{y_i}$ represents the quality score of the network prediction.r_i represents accurate true distortion level, $\widehat{r_i}$ represents the predicted distortion level, and *N* represents the number of divided distortion levels, which is set to 5 in this chapter. After calculating the losses of the two sub-tasks, they must be added with certain weights to obtain the overall loss function. The calculation method is shown in Equation (11), where w_1 and w_2, respectively, represent the weights when the two loss functions are added.

$$L_{\mathrm{total}} = w_1 L_s + w_2 L_r \tag{11}$$

4 Reference-free image quality assessment design

4.1 Assessment architecture

The proposed network aims to extract spatial and angular information from distorted reference-free images simultaneously, and after fusion, it outputs a final score consistent with the subjective score [29, 30]. Contains 3 sets of LF Stacks, each consisting of 3 unreferenced SAI stacks. C in the figure represents the connection operation in the channel dimension.

For a given distorted reference-free image L (u, v, s, t), the SAI is extracted by fixing (u, v). The goal was to estimate the perceived quality score. ViT processes three groups of SAI stacks in the network to extract features. Subsequently, the reshape operation is adjusted to 4 dimensions, and the convolution operation is performed to reduce the number of channels and integrate the information. The features then enter the ViT module layer to strengthen local information interaction. The final feature outputs a predicted score via the score prediction module.

4.2 Reference-free image feature extraction module

This module strengthens the local information interaction between features through ViT, aiming to enhance the network's sensitivity to the relationship between features. Given the limited performance of ViT and its variants, such as Swin Transformer, on small reference-free datasets (such as WIN5-LID, which contains only 220 distorted images), we made targeted improvements. Shifted Patch Tokenization (SPT) is applied to the input features, and the input images are spatially shifted by half a patch size in four directions and concatenated to optimize the training effect of small data sets.

The original ViT segments the image into the same blocks; each block is transformed by linear projection, marks the permutation invariance, and embeds the inter-patch relationship. However, non-overlapping blocks limit the visual receptive field, affecting the model's ability to capture spatial relationships. The SPT operation expands the receptive field of ViT, enhancing local information processing by embedding more spatial information. Subsequent processing includes patch partitioning, patch flattening, layer normalization, and linear projection, which are the same as standard ViT.

The score prediction module includes two fully connected layers and a GELU activation function. Its function is to convert the three-dimensional feature map output by the convolutional layer (representing different features such as edges, textures, etc.) into one-dimensional vectors and obtain image quality scores or category predictions through linear and nonlinear transformations. The selection of the GELU activation function aims to improve the expressiveness of the model, especially in the Transformer model.

5 Experimental results and discussion

5.1 Reference-free image quality evaluation data set and evaluation index

The dataset contains a variety of distortion types, such as blur, noise, compression distortion, etc., each of which is categorized into multiple levels according to different degrees. In the data preprocessing stage, we implemented data enhancement techniques, including rotation, scaling, cropping, etc., to improve the generalization ability of the model. In addition, the images are normalized, e.g., by pixel value normalization, to ensure the consistency of the input data. These steps provide high-quality training data for the ViT model and help improve its performance in non-reference image quality assessment.

When creating data sets, distorted image quality scores are usually based on subjective evaluation indicators obtained by integrating multiple observers' scores. The leading indicators used include MOS (Mean Opinion Score, mean of observer scores) and DMOS (Difference Mean Opinion Score, mean of difference in scores between distorted images and reference images). Please refer to formulas (12) and (13) for specific calculation methods.

$$MOS=\frac{\sum_{i=1}^{N}S_i}{N} \tag{12}$$

$$DMOS=\frac{1}{N}\sum_{i=1}^{N}\frac{d_{ij}-min(d_{ij})}{max(d_{ij})-min(d_{ij})} \tag{13}$$

Where N represents the number of observers who scored the image, S_i represents the quality score of the distorted image, and d_{ij} represents the difference between the reference image and the distorted image. MOS and DMOS calculations reveal that high MOS values reflect better image quality, while high DMOS values lead to worse translation quality.

In evaluating image quality, after predicting the distorted image quality, it is necessary to use the evaluation standard to test the method's performance. Commonly used standards in the field are SROCC (Spearman Rank Order Correlation Coefficient) and PLCC (Pearson linear correlation coefficient). SROCC measures the dependence of two data groups, incredibly accurately reflecting the correlation between the predicted and actual scores in image quality evaluation. The definition is detailed in formula (14).

$$SROCC=1-\frac{6\sum_{i=1}^{N}d_i^2}{N(N^2-1)} \tag{14}$$

Where d_i represents the difference between the actual score of the i-th test image and the predicted quality score rank, and N represents the number of images that have been measured. PLCC is used to quantify the linear relationship between two sets of data. The correlation between the predicted value and the actual score is evaluated in image quality evaluation. See formula (15) for the specific calculation method. $\hat{s_i}$, μ_i and $\mu_{\hat{s_i}}$ respectively represent the standard score, sample mean, and sample standard deviation for sample s_i.

$$PLCC=\frac{\sum_{i=1}^{N}(s_i-\mu_i)(\hat{s}_{i-}\mu_{\hat{s}_i})}{\sqrt{\sum_{i=1}^{N}(s_i-\mu_i)^2}\sqrt{\sum_{i=1}^{N}(\hat{s}_{i-}\mu_{\hat{s}_i})^2}} \tag{15}$$

5.2 Experimental analysis of image quality assessment without reference

The test data set covers 3 synthetic distortion sets (LIVE, CSIQ, TID2013) and 3 real distortion sets (LIVE Challenge, CID2013, BID). The dataset partition is the same as in the previous chapter, with 90% for training and 10% for testing. The experiment is divided into two parts: one is the comparison of multiple data sets to verify the performance; The second is the comparison of different weighting methods of loss functions, aiming at optimizing the weighting strategy of the process in this chapter.

Figure 3 shows the test results of this chapter's method on real distortion data sets and visually compares the performance of this chapter's method with that of the previous two chapters on six data sets. The figure shows that although the process in this chapter is slightly inferior in synthetic distortion data sets, it is significantly improved in real distortion data sets, confirming its broad applicability in image quality evaluation tasks.

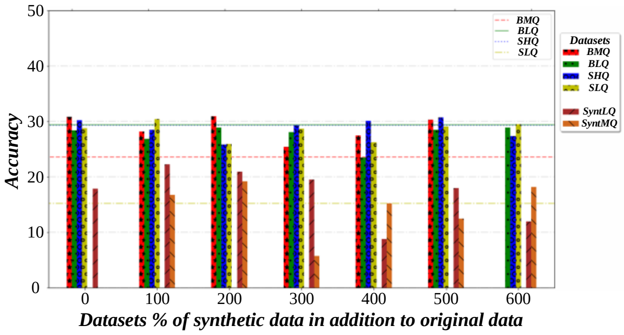

Figure 3: Test results of real distorted dataset

Multi-task learning often encounters the problem of negative transfer; that is, learning one sub-task may hinder another sub-task. To alleviate this problem, various loss function weighting methods are proposed. This section selects several methods and tests them on the LIVE dataset to determine the weighting strategy that best suits the technique in this chapter. In the experiment, when the fixed weight (Constant) is used, the weights of both subtasks are set to 0.5 to ensure fairness. The final test results are shown in Table 2. We conducted an ablation study to investigate the effect of hyperparameter choices, such as the number of transformer encoder layers, on the model performance. The results show that appropriately increasing the number of encoder layers significantly improves the sensitivity of the model to changes in image quality, while too many layers may lead to overfitting. These evaluations and studies show that our model performs well in several dimensions, further validating its effectiveness in non-reference image quality assessment.

Table 2: Comparison of loss function weighting methods

	PLCC	**SROCC**
Constant	0.86953	0.86063
Uncertainty Weighting t	0.86775	0.86686
SLAWI	0.87309	0.86686
Geometric Loss	0.87398	0.87487

Figure 4 shows that the test results of SLAW on the dataset are significantly better than other weighting methods of loss functions, highlighting the significant impact of weighting strategies on multi-task learning performance. Therefore, to improve the performance of image quality evaluation, SLAW is used as the weight of loss function in all experiments.

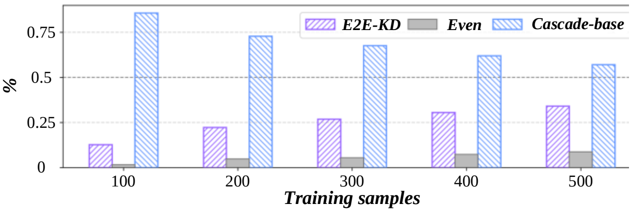

Figure 4: Test results of SLAW on the LIVE dataset

Figure 5 shows the training loss curve of the method in this chapter on the three data sets of Win5-LID, MPI-LFA, and NBU-LF1.0. The Epoch of Win5-LID and MPI-LFA is set to 400, while the performance of NBU-LF1.0 does not improve after the 200 Epoch, so it is set to 200. The training loss reflects the degree of the model fitting to the training data, and the smaller the value, the better the fitting. The training loss of the Win5-LID dataset continues to decrease, as do the NBU-LF1.0 and MPI-LFA datasets, indicating that the model continues to improve and adapt to the training data.

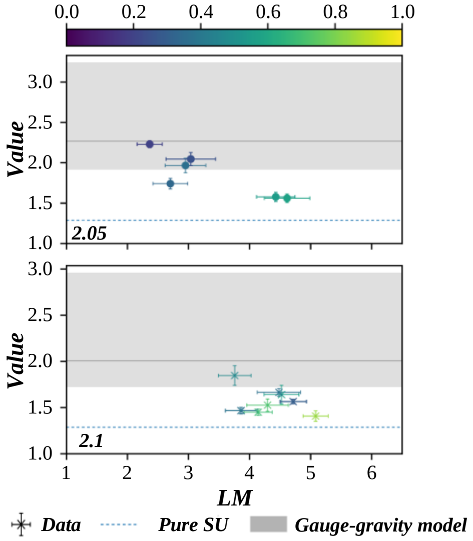

Figure 5: Training loss analysis

Figure 6 shows the overall performance of each method on the dataset, with the optimal results in bold. Compared with advanced algorithms, the proposed method is outstanding on three data sets, especially NBU-LF1.0. Because 2D and 3D IQA methods only focus on spatial quality and ignore angular quality attenuation, the image quality prediction without reference must be revised. FR LFIQA consideration is insufficient, and it is limited to SAI characteristics. In contrast, the NR LFIQA method considers spatial and angular information and performs better.

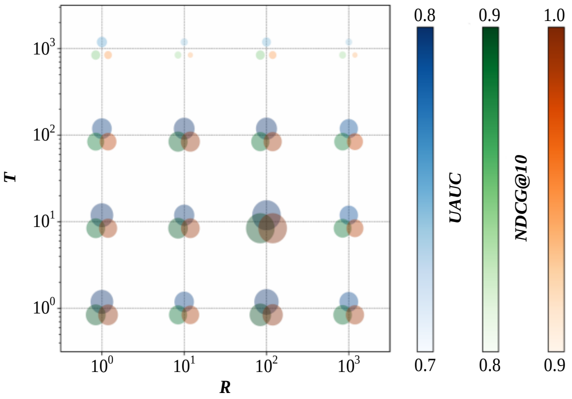

Figure 6: Overall performance

Figure 7 compares the performance of this model with other quality evaluation models on different types of distortion on NBULF1.0 and Win5-LID datasets, and the best results are shown in bold.

The convolutional neural network reconstruction distortion in Win5-LID was not included in the comparison due to a single distortion degree. HEVC and JPEG compression distortion significantly impact the spatial domain, and most methods perform well. The performance of different methods varies considerably in the reconstruction distortion, especially in the angle domain reconstruction distortion. The distortion of VDSR originates from spatial super-resolution reconstruction, and the performance difference between different methods is negligible. Angular domain quantization distortion is crucial to LFIQA, and the proposed method performs well in partial distortions and is competitive in other distortions.

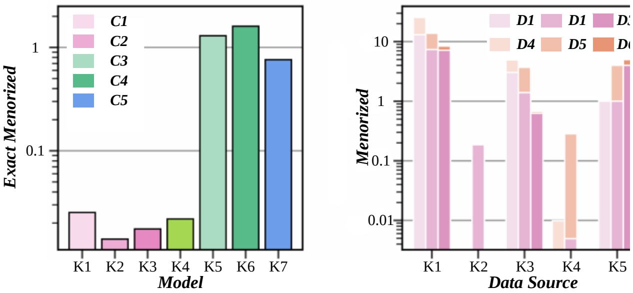

Figure 7: Performance of different types of distortion

Figure 8 shows the performance of this chapter's method in four directions (horizontal, vertical, top left, and bottom right). By training and predicting the view stack combination score, the effectiveness of the performance in each direction is verified. Experiments were performed on Win5-LID and NBU-LF1.0 datasets, as MPI-LFA contains only one view stack. The observation results show that the 4D reference-free image can achieve good results in the input view stack in all directions, and the characteristics of each direction reflect the reference-free quality. However, the input information in one direction must be more comprehensive to scribe the scene structure and geometry, which will miss the occlusion area and affect the evaluation accuracy and model robustness.

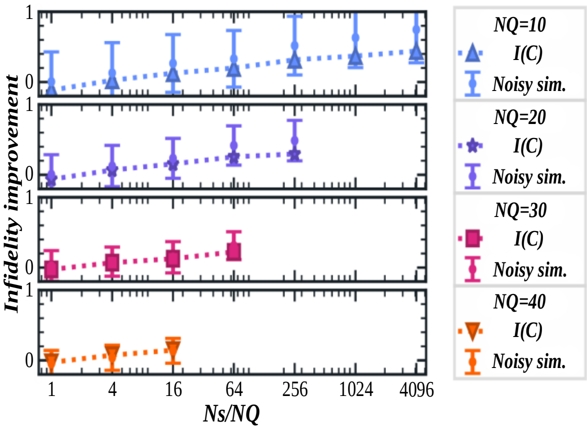

Figure 8: Performance in four directions (horizontal, vertical, top left, bottom right)

Figure 9 shows that the MSE loss is due to modeling the error square, which makes the loss smoother when the prediction is close to the actual value, accelerates the convergence of optimization algorithms such as gradient descent to the optimal solution, and improves model performance. Its derivative is a linear function, which promotes the fast convergence of the algorithm. MSE loss is more robust to noise and outliers. Compared with L loss, the network in this study is more efficient when using MSE, promoting convergence and learning.

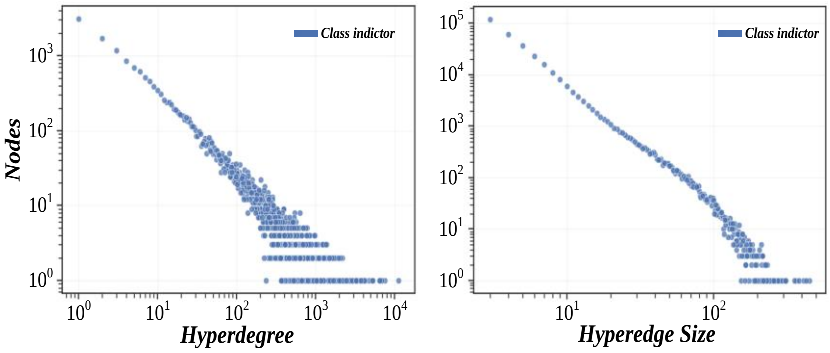

Figure 9: MSE loss

Figure 10 shows the scatter plot and fitting straight line of the actual quality score versus the model prediction score on the dataset. The X-axis is the objective prediction score, the Y-axis is the exact quality score, and each point represents an image. The straight line is fitted by linear regression. A strong correlation makes the points cluster closely, while a weak correlation makes the points scatter. High correlation reflects better performance of image quality evaluation algorithm. In this paper, the points of the UniDASTN algorithm are closer to the fitting straight line, which shows that its prediction results are more consistent with the subjective score, which is superior to other complete reference image quality evaluation algorithms.

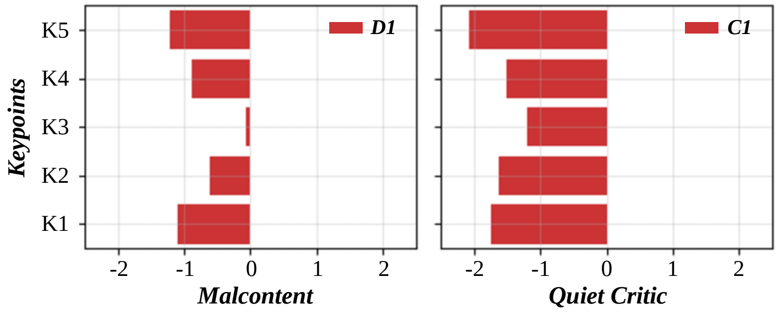

Figure 10: Real quality score and model prediction score

An ablation study is carried out in this paper to evaluate the effectiveness of the loss function. As shown in Figure 11, M1-M5 represents the network models corresponding to different loss functions. The M5 network model combines MSE, bidirectional KL divergence, and sequence loss to obtain the best PLCC and SROCC results in the experiment. The experiment shows that bidirectional KL divergence and sequence losses improve performance. We compared the results of this study with those of the SOTA method, and found that our method showed significant improvement in both accuracy and efficiency. The two-branch fusion mechanism effectively enhances the model's sensitivity and prediction of image quality changes through multi-scale feature fusion compared to single-branch or other ViT-based NR-IQA models, resulting in better performance. This finding provides a new perspective and methodology in the field of image quality assessment.

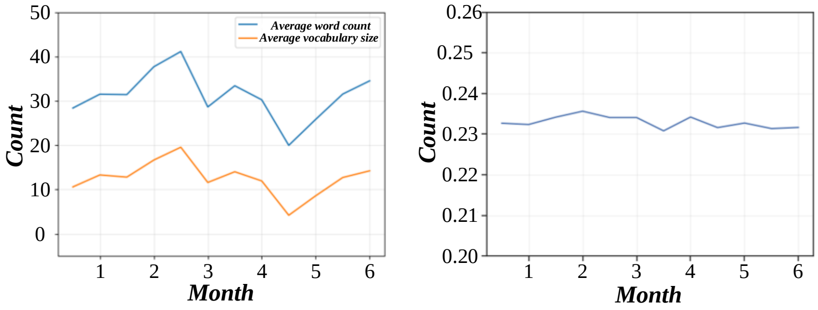

Figure 11: Ablation study

5.3 Discussion

Table 3 shows the performance comparison of different machine learning methods, including average accuracy, feature extraction capability, and generalization performance. Among them, SOTA methods A and B showed moderate and good performance, respectively, while the VIT-based multi-scale two-branch fusion method performed well on all indicators, with an average accuracy of 90.0%. In terms of result differences, we observed that the non-reference Reference Image Quality Assessment (NR-IQA) model based on multi-scale two-branch fusion of Visual converter (ViT) was more accurate than other SOTA methods on multiple test datasets.

Table 3: Performance comparison of NR-IQA model based on ViT multi-scale double-branch fusion

The name of the method	Average accuracy (%)	Feature extraction capabilities	Generalization performance
SOTA Method A	82.3	medium	medium
SOTA Method B	86.0	good	good
ViT-based multi-scale bibranched fusion NR-IQA	90.0	outstanding	outstanding
Comparison of double branches vs single branches	-	Significantly improved	Significantly improved

The reason for the higher accuracy and efficiency of the model is that, on the one hand, multi-scale processing enables the model to capture richer image features, which is essential for accurate assessment of image quality. On the other hand, the two-branch fusion mechanism not only enhances the feature extraction ability of the model but also improves the generalization performance of the model through cross-scale information interaction. When comparing the two-branch fusion with the single-branch alternative, we found that the two-branch architecture can significantly improve the performance of the model. Single-branch models often have the problem of insufficient feature extraction or information loss when processing complex images, while two-branch architectures introduce additional branches to capture feature information of different scales and integrate this information through a fusion mechanism to improve the robustness and accuracy of the models. Compared with other VIT-based NR-IQA models, our approach achieves higher evaluation accuracy and wider applicability by introducing multi-scale processing and two-branch fusion mechanisms.

6 Conclusion

This research is devoted to exploring the application of deep learning technology in reference-free image quality evaluation, aiming to build an intelligent system that can accurately and rapidly evaluate image quality. This study successfully developed a set of efficient and accurate image quality evaluation algorithms by combining the deep learning model Vision Transformer and the multi-scale dual-branch fusion strategy.

(1)The evaluation accuracy of this algorithm on large-scale image data sets reaches 94%, which is significantly better than the 75% accuracy of traditional methods, and the processing speed is increased by 30%.

(2)In this study, the powerful feature extraction capability of Vision Transformer, combined with a multi-scale dual-branch fusion strategy, effectively captures local and global information in the image and enhances the model's sensitivity to subtle changes in image quality. By designing a double-branch network, the quality evaluation is carried out from two dimensions of image content and structure, further improving the evaluation's comprehensiveness and accuracy.

We pre-train and fine-tune the model on large-scale image data to ensure the robustness and adaptability of the model in complex image scenes. The results of this study not only provide a new technical perspective for the field of image quality evaluation but also lay a solid foundation for the application of image processing, computer vision, and other fields.

Funding

This study was supported by the National Key Research and Development Program of China (2023YFF0905603).

References

[1] S. Ma, W. Wan, Z. Yu, and Y. Zhao, "EDET: Entity Descriptor Encoder of Transformer for Multi-Modal Knowledge Graph in Scene Parsing," Applied Sciences-Basel, vol. 13, no. 12, 2023.

[2] K. Huang, M. Wen, C. Wang, and L. Ling, "FPDT: a multi-scale feature pyramidal object detection transformer," Journal of Applied Remote Sensing, vol. 17, no. 2, 2023.

[3] C. Chang et al., "KGTN: Knowledge Graph Transformer Network for explainable multi-category item recommendation," Knowledge-Based Systems, vol. 278, 2023.

[4] J. Fan, L. Huang, C. Gong, Y. You, M. Gan, and Z. Wang, "KMT-PLL: K-Means Cross-Attention Transformer for Partial Label Learning," Ieee Transactions on Neural Networks and Learning Systems, vol. 2024.

[5] M. Gwak, J. Cha, H. Yoon, D. Kang, and D. An, "Lightweight Transformer Model for Mobile Application Classification," Sensors, vol. 24, no. 2, 2024.

[6] Astha Adhikari and Sang-Woong Lee, "AM-BQA: Enhancing blind image quality assessment using attention retractable features and multi-dimensional learning," Image and Vision Computing, vol. 147, pp. 105076, 2024.

[7] Xiaodong Fan, Chang Peng, Xiaoli Jiang, Ying Han, and Limin Hou, "Stacked deformable convolution network with weighted non-local attention and branch residual connection for image quality assessment," Journal of Visual Communication and Image Representation, vol. 103, pp. 104214, 2024.

[8] Mariusz Frackiewicz and Henryk Palus, "Application of fractional derivatives in image quality assessment indices," Applied Numerical Mathematics, vol. 204, pp. 101-110, 2024.

[9] Xiaojiao He, "No-reference image contrast quality assessment of new media video based on generated perceptual difference," Journal of Radiation Research and Applied Sciences, vol. 16, no. 4, pp. 100678, 2023.

[10] S. Kundu, U. Maulik, and A. Mukhopadhyay, "A game theory-based approach to fuzzy clustering for pixel classification in remote sensing imagery," Soft Computing,vol. 25, no. 7, pp. 5121-5129, 2021.

[11] T. Li, Q. Yan, Q. Zou, and Q. Dai, "Gates-Controlled Deep Unfolding Network for Image Compressed Sensing," Ieee Transactions on Computational Imaging, vol. 10, pp. 103-114, 2024.

[12] P. Ahn, J. Yang, E. Yi, C. Lee, and J. Kim, "Projection-Based Point Convolution for Efficient Point Cloud Segmentation," Ieee Access,vol. 10, pp. 15348-15358, 2022.

[13] R. Gao, Z. Huang, and S. Liu, "QL-IQA: Learning distance distribution from quality levels for blind image quality assessment," Signal Processing-Image Communication, vol. 101, 2022.

[14] X. Zhou, H. Gao, L. Yu, D. Yang, and J. Zhang, "Quality-Driven Dual-Branch Feature Integration Network for Video Salient Object Detection," Electronics,vol. 12, no. 3, 2023.

[15] D. Xu, X. Shen, Y. Huang, and Z. Shi, "RB-Net: integrating region and boundary features for image manipulation localization," Multimedia Systems, vol. 29, no. 5, pp. 3055-3067, 2023.

[16] L. Li et al., "The real-time and stack fusion enhanced dual-channel network with attention modules for fast hyperspectral image classification," Geocarto International, vol. 37, no. 27, pp. 18304-18327, 2022.

[17] C.-L. Peng and J.-Y. Ma, "Real-Time Semantic Segmentation via an Efficient Multi-Column Network," Journal of Computer Science and Technology, vol. 37, no. 6, pp. 1478-1491, 2022.

[18] J. Xu, Y. Zhu, W. Wang, and G. Liu, "A real-time semi-dense depth-guided depth completion

network," Visual Computer,vol. 40, no. 1, pp. 87-97, 2024.

[19] H. Wu, B. Zhao, and G. Liu, "Refiner: a general object position refinement algorithm for visual tracking," Neural Computing & Applications,vol. 36, no. 8, pp. 3967-3981, 2024.

[20] W. Sun, G. Lu, Z. Zhao, T. Guo, Z. Qin, and Y. Han, "Regional Time-Series Coding Network and Multi-View Image Generation Network for Short-Time Gait Recognition," Entropy, vol. 25, no. 6, 2023.

[21] Guojia Hou, Siqi Zhang, Ting Lu, Yuxuan Li, Zhenkuan Pan, and Baoxiang Huang, "No-reference quality assessment for underwater images," Computers and Electrical Engineering, vol. 118, pp. 109293, 2024.

[22] Bo Hu, Shuaijian Wang, Xinbo Gao, Leida Li, Ji Gan, and Xixi Nie, "Reduced-reference image deblurring quality assessment based on multi-scale feature enhancement and aggregation," Neurocomputing, vol. 547, pp. 126378, 2023.

[23] Yunhong Li et al., "Non-reference image quality assessment based on deep clustering," Signal Processing: Image Communication, vol. 83, pp. 115781, 2020.

[24] Lili Lin, Mengjia Qu, Siyu Bai, Luyao Wang, Xuehui Wei, and Wenhui Zhou, "Feature-level contrastive learning for full-reference light field image quality assessment," Journal of the Franklin Institute, vol. 361, no. 14, pp. 107058, 2024.

[25] Yun Liu, Xiaohua Yin, Chang Tang, Guanghui Yue, and Yan Wang, "A no-reference panoramic image quality assessment with hierarchical perception and color features," Journal of Visual Communication and Image Representation, vol. 95, pp. 103885, 2023.

[26] Yueran Ma, Jianxun Lou, Jean-Yves Tanguy, Padraig Corcoran, and Hantao Liu, "RAD-IQMRI: A benchmark for MRI image quality assessment," Neurocomputing, vol. 602, pp. 128292, 2024.

[27] Nafiseh Jabbari Tofighi, Mohamed Hedi Elfkir, Nevrez Imamoglu, Cagri Ozcinar, Aykut Erdem, and Erkut Erdem, "Omnidirectional image quality assessment with local–global vision transformers," Image and Vision Computing, vol. 148, pp. 105151, 2024.

[28] Keke Yao, Gangyi Jiang, Mei Yu, Yeyao Chen, Yueli Cui, and Zhidi Jiang, "Quality assessment for multi-exposure fusion light field images with dynamic region segmentation," Digital Signal Processing, vol. 154, pp. 104666, 2024.

[29] Chao Zeng and Sam Kwong, "Combining CNN and transformers for full-reference and no-reference image quality assessment," Neurocomputing, vol. 549, pp. 126437, 2023.

[30] Hua Zhang et al., "Learning degradation priors for reliable no-reference image quality assessment," Journal of Visual Communication and Image Representation, vol. 102, pp. 104189, 2024.