IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 40, No. 6, December, 2023, pp. 2553-2563
Journal homepage: http://iieta.org/journals/ts

Analyzing and Optimizing Virtual Reality Classroom Scenarios: A Deep Learning Approach

Lin Jiang[1*], Xiao Lu[2]

[1] School of Foreign Languages, Shenyang University of Technology, Shenyang 110178, China
[2] School of Marxism, Shenyang University of Technology, Shenyang 110178, China

Corresponding Author Email: 12282@sut.edu.cn

https://doi.org/10.18280/ts.400618

Received: 12 June 2023
Revised: 19 October 2023
Accepted: 25 October 2023
Available online: 30 December 2023

Keywords:
Virtual Reality (VR) classroom, deep learning, scene analysis, feature enhancement, feature distillation, multi-scale information, transformer, attention mechanism, semantic segmentation, classification optimization

ABSTRACT

With the continuous progress of information technology, Virtual Reality (VR) technology has been more and more widely used in the field of education, and the VR classroom has revolutionized the traditional education model due to its unique interactivity and immersion. However, the complex and changing VR classroom scenes bring challenges to effective scene analysis and optimization. Although existing deep learning methods have made significant progress in image processing, they still face the problems of capturing insufficient detail information and under-utilizing global information when accurately segmenting and classifying VR classroom scenes. To address these problems, this study proposes a series of innovative approaches. The first part investigates VR classroom scene segmentation based on feature enhancement and feature distillation. By designing an attention mechanism with multi-pooling compression incentives and a feature dehazing branch structure with "enhance-refine-subtract" strategies, the network's ability to extract valid information is significantly improved and the interference of invalid information is effectively reduced, which greatly enhances the accuracy of semantic segmentation. The second chapter talks about the optimization of VR classroom scene classification based on multi-scale global information enhancement. By incorporating the *Transformer* structure, multi-scale information is extracted effectively, global associated information is utilized comprehensively, information processing mechanism in the classification process is optimized, and classification performance is enhanced. Results attained in this study not only improves our understanding of VR classroom scenes, but also provides new insights and technical approaches for the application of deep learning models in processing complicated scenes. Moreover, findings of this paper could portend far-reaching implications in the fields of educational technology and computer vision, and broaden the application range of VR classroom.

1. INTRODUCTION

Today, as the deep learning technology in the field of computer vision is developing at a fast speed, the new idea of VR classroom has attracted widespread attention, and now it has been adopted as an emerging education method [1-3]. In this context, some researchers have included the analysis of VR classroom scenes as a research topic of theirs [4, 5], and they have realized that by accurately analyzing and understanding the visual content in a VR classroom, the quality of teaching and the immersion of learners can be greatly enhanced [6-8]. However, the complexity and variability of VR classroom scenes, as well as the noise, and the interference factors in the environment, can all bring challenges to scene analysis and optimization.

The importance of VR classroom scene analysis lies in not only the construction quality of the virtual teaching environment, but also the effective use of educational resources, and the innovation of teaching methods [9]. If we can accurately segment and classify these scenes through deep learning models, then more intuitive teaching assistance can be provided to teachers, and more personalized learning experiences could be provided to students [10-14]. Besides, optimized scene analysis techniques can be extended to other VR applications, such as games and simulation training, and these can largely broaden the application range of VR.

Although scholars in the field have proposed many scene analysis methods so far, yet these methods still have certain defects and shortcomings when dealing with VR classroom scenes [15, 16]. For example, during scene segmentation, common deep learning methods tend to ignore the minute details and boundary information in the environment, and these can result in inaccurate segmentation results [17, 18]. Meanwhile, the extraction and utilization of global information in scene classification is insufficient, making the classification results unable to fully reflect the details and deep-level features of the scene, and these problems limit the accuracy and reliability of VR classroom scene analysis [19-22].

The main content of this paper focuses on two core research parts: the first one is VR classroom scene segmentation based on feature enhancement and feature distillation. This study

proposes a new attention mechanism with multi-pooling compression incentives, and a feature dehazing branching structure with "enhance-refine-subtract" modules. These innovative approaches significantly enhance the network's ability to capture valid information and reduce the interference of invalid information, thus achieving significant improvements in the performance of semantic segmentation of scene images. Second, this paper investigates the optimization of VR classroom scene classification based on multi-scale global information enhancement. The introduction of *Transformer* architecture allows for effective extraction of multi-scale spatial information and deep integration of global data. These findings not only significantly optimize the understanding and representation of VR classroom scenes but also provide new perspectives and methodologies for the application of deep learning in complex scene analysis, holding significant theoretical and practical value.

2. SCENE SEGMENTATION IN VR CLASSROOMS BASED ON FEATURE ENHANCEMENT AND DISTILLATION

In VR classroom environments, scenes often contain a variety of complex visual elements, such as diverse teaching materials, simulated objects, and learner interactions. These elements can easily lead to information confusion and loss of details in traditional segmentation methods. Particularly under conditions of changing lighting, perspective shifts, and instances of blurring or occlusion in the scene, effective information extraction becomes notably challenging. In response to these practical issues in the analysis and optimization of VR classroom scenes, this paper introduces an attention method based on multi-pooling compression excitation. The multi-pooling structure enhances the network's ability to capture features of varying scales. The compression excitation mechanism, through weighted allocation, intensifies focus on pertinent features while suppressing irrelevant information, thus improving the precision in recognizing educational content and learner interactions during the segmentation task. Moreover, this method demonstrates superiority in removing or reducing visual noise that may stem from VR technology itself, ensuring that the model maintains high segmentation performance under less than ideal visual conditions. This implies that the method proposed in this paper maintains stable segmentation effects, whether in dimly lit scenes or against complex backgrounds, providing more accurate and robust scene analysis capabilities for VR classrooms.

Figure 1 illustrates the structure of the multi-pooling compression excitation module. The core design of the proposed module lies in the parallel use of global max pooling and global average pooling layers, which capture different types of global information. The global max pooling layer highlights the most significant features, namely the salient signals in the scene, while the global average pooling layer provides statistical information of the feature mappings, reflecting the overall distribution. These two types of information are then dimensionally reduced through their respective fully connected layers, compressing the features and reducing the number of parameters, thereby also enhancing the model's generalization ability. Subsequently, the *ReLU* activation layer is employed to introduce non-linearity, enhancing the network's capability to express and learn more complex non-linear relationships between features. Following this, the features output from the aforementioned two paths are further encoded through individual *FC* layers, making the features more compact while retaining necessary information. Then, these two features are fused through an addition operation, effectively combining the global information extracted by the max and average values, offering a more comprehensive perspective of the features. Finally, the fused features are activated through a *Sigmoid* activation function, yielding weights between 0 and 1, which represent the importance of each channel. These weights, when multiplied channel-wise with the input features, enable recalibration of the original features, thereby emphasizing beneficial information and suppressing insignificant signals.

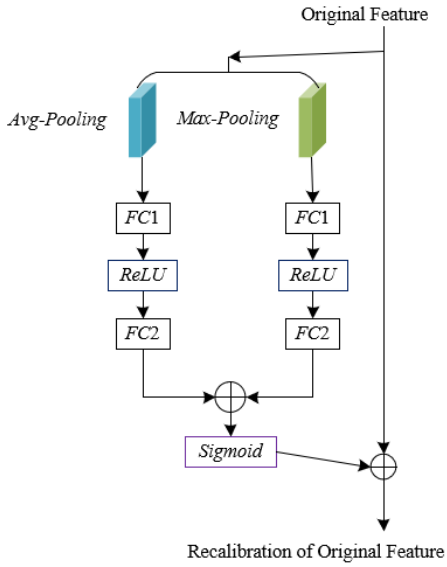

Figure 1. Structure of the multi-pooling compression excitation module

Initially, the input feature map is processed through the global max pooling and global average pooling layers, capturing the salient and overall statistical information of the scene's features, respectively. Subsequently, these global pieces of information are compressed through their respective fully connected layers, aiming to reduce dimensions, thereby decreasing the model's parameters and computational load, and preparing compact feature representations for the subsequent excitation step. Assuming the compressed features are represented by D_{MS} and D_{AS}, the number of channels in feature D is denoted by Z, and the height and width of the features are represented by G and Q, the two-dimensional input feature in the z-th channel, and the compressed features post-global max pooling and global average pooling are denoted by D^z, D^z_{MS}, and D^z_{AS}, respectively. The computational formulas are as follows:

$$D^z_{MS} = \underset{u\in G, k\in Q}{MAX}\left(D^z(u,k)\right) \tag{1}$$

$$D^z_{AS} = \frac{1}{G\times Q}\sum_{u=0}^{G}\sum_{k=0}^{Q} D^z(u,k) \tag{2}$$

The compressed features D^z_{MS} and D^z_{AS} are passed through a *ReLU* activation layer to introduce non-linearity, enabling the model to capture more complex feature relationships.

Subsequently, these features are further encoded through a fully connected layer, generating an independent weight value for each feature channel. Suppose the weight matrices for DZ_1 and DZ_2 are represented by $Q_1 \in R^{Z*Z/e}$ and $Q_2 \in R^{Z*Z/e}$, respectively. The *ReLU* activation function is denoted by σ, and the process is expressed as follows:

$$D'_{MEX} = Q_2\sigma\left(Q_1 D_{MS}\right) \tag{3}$$

$$D'_{AEX} = Q_2\sigma\left(Q_1 L D_{AS}\right) \tag{4}$$

The two sets of channel weights obtained in the previous step, D'_{MEX} and D'_{AEX}, are added together to yield the aggregated feature D'_{EX}, integrating the global information captured by both maximization and average pooling strategies. The aggregated feature is then processed through a *Sigmoid* activation function, outputting the final weights for each channel. These weights, ranging between 0 and 1, represent the relative importance of each channel to the task. Assuming the *Sigmoid* activation function is denoted by σ, the final expression for the excited feature D'_{EX} is as follows:

$$D_{EX} = \delta\left(D'_{EX}\right) = \delta\left(D'_{MEX} + D'_{AEX}\right) \tag{5}$$

Suppose the input variable value is represented by c, and the final function output value is denoted by $\delta(c)$. The specific formula for the *Sigmoid* activation function is:

$$\delta(c) = \frac{1}{1-e^{-c}} \tag{6}$$

Finally, the weights output by the *Sigmoid* function are multiplied channel-wise with the original input feature map, achieving feature recalibration. This step enhances focus on useful features while suppressing unimportant or interfering information, thereby maintaining the original feature structure while highlighting the expression of key information in the analysis of VR classroom scenes. Assuming the excited one-dimensional feature is represented by D_{EX}, and the recalibrated feature is denoted by $\tilde{D}$, which is the product of D_{EX} and the original feature D in the corresponding channels, the expression is as follows:

$$\tilde{D} = D_{EX} \cdot D \tag{7}$$

Figure 2 displays the overall network structure that incorporates the multi-pooling compression excitation module. This paper provides solutions to a series of practical problems encountered in the analysis and optimization of VR classroom scenes by introducing a feature dehazing branch structure with an "enhance-refine-subtract" enhancement module, along with a feature distillation module incorporating an attention mechanism. Initially, the dehazing branch structure effectively addresses the issue of visual information loss in VR environments caused by simulated fog effects. The "enhance" step improves the contrast and clarity of features, thereby enhancing their recognizability; the "refine" step further precisely adjusts these features, ensuring the preservation of details; the "subtract" step effectively removes the scattering and color degradation effects caused by the fog, restoring the essential characteristics of the scene. Subsequently, the integrated feature distillation module, through its attention mechanism, further focuses and optimizes key features, ensuring effective integration of features in multi-task learning and enhancing the model's ability to parse complex scenes in semantic segmentation tasks. Figure 3 demonstrates the backbone network structure incorporating the feature dehazing branch and the feature distillation module.

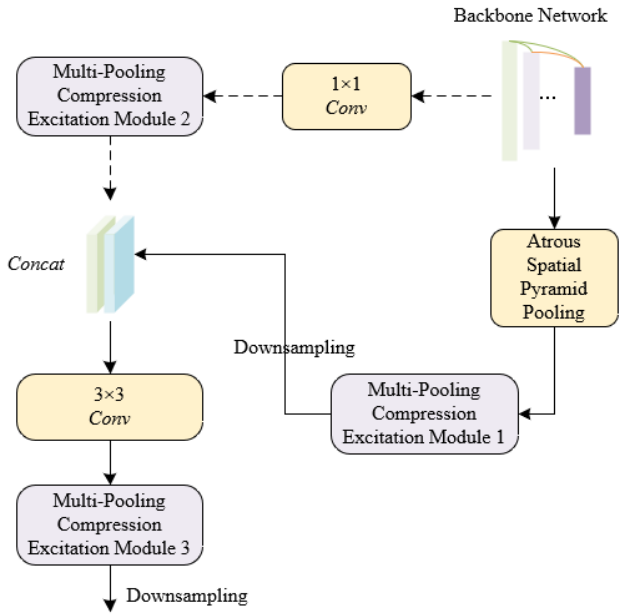

Figure 2. Network structure incorporating multi-pooling compression excitation module

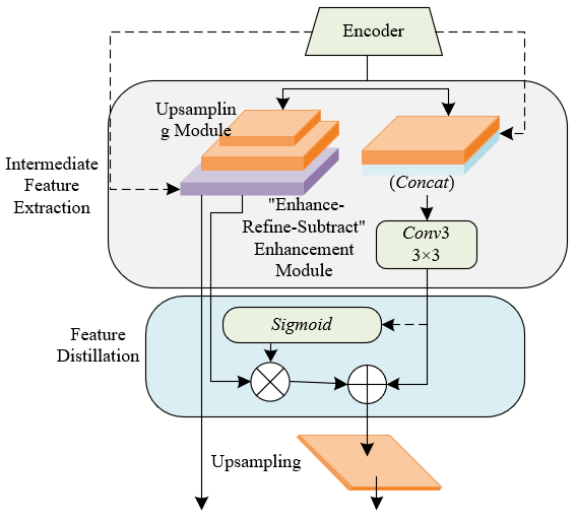

Figure 3. Backbone network incorporating feature dehazing branch and feature distillation module

The feature dehazing branch structure first upsamples deep features through two stride-2 transposed convolution modules. This process not only increases the spatial resolution of the feature map but also aligns the number of feature channels with corresponding stages in the decoder, laying the foundation for feature fusion. Subsequently, these upsampled features are fused with the jump connection features from the encoder. This step utilizes the rich edge and texture information contained in shallow features to complement deep features, thereby obtaining a more comprehensive feature representation. Further, the fused features are input into the "Enhance-Refine-Subtract" enhancement module, the core

function of which is to improve the quality and distinction of features. The output fog-free features are significantly enhanced in quality through structured processing in the "enhance," "refine," and "subtract" components. Assuming the enhanced image after v iterations is represented by $\hat{K}_v$, the image deblurring method by $h(.)$, and the input blurred image by U, the following formula represents the expression of the enhancement module:

$$\hat{K}_{v+1} = h\left(U + \hat{K}_v\right) - \hat{K}_v \tag{8}$$

Assuming the enhanced feature at stage v is represented by k_v, the shallow feature obtained from the encoder at stage v by u_v, the 2x upsampling operation by $\uparrow_2$, and the feature refinement operation unit by $H(.)$, the module can also be represented as:

$$k_v = H\left(u_v + \left(k_{v+1}\right)\uparrow_2\right) - \left(k_{v+1}\right)\uparrow_2 \tag{9}$$

The feature distillation module with an attention mechanism introduced in this paper is an efficient multi-task learning strategy. It enhances the performance of the final task by synthesizing feature information generated from different tasks. This module's structural feature lies in its ability to operate within a multi-dimensional feature space, concatenating the fog-free features from the feature dehazing branch with the refined segmentation features from the decoder. In this process, the attention mechanism plays a key role by learning the importance distribution of these features, selectively strengthening information beneficial to the final task while suppressing irrelevant or interfering information. In the feature distillation module, the specific role of the attention mechanism is reflected in its weight distribution among different features. These weights are not randomly assigned but learned through the network, reflecting the contribution of each feature to the final task. The attention mechanism dynamically adjusts the information flow between features, enabling the network to focus more on information useful for tasks like analyzing VR classroom scenes.

Suppose the intermediate feature of the target task is represented by D^u_1, the attention feature map by T^u, the intermediate feature of the auxiliary task by D^u_2, and the distilled feature by D^p. Multiplying D^u_2 with T^u at the element level to obtain the weighted feature and adding it to D^u_1 yields D^p. The following formula represents the feature distillation module process expression:

$$D^p = D^u_1 + T^u \cdot D^u_2 = D^u_1 + \delta\left(QD^u_1\right) \cdot D^u_2 \tag{10}$$

3. OPTIMIZATION OF VR CLASSROOM SCENE CLASSIFICATION BASED ON MULTI-SCALE GLOBAL INFORMATION ENHANCEMENT

In a complex and variable virtual environment, accurately recognizing and classifying different teaching elements and interactive behaviors presents a challenge for researchers in related fields. Conventional classification methods generally can not fully take into account the global information of different scales, so their classification performance is often limited when dealing with large structures and detailed features. For this reason, this paper proposes an optimization method for VR classroom scene classification based on multi-scale global information enhancement. The method optimizes the capture of global information and the integration of features by introducing a multi-scale global information enhancement module, which ensures that the model is able to capture the global characteristics of the large-scale scene layout and teaching interactions, and also focuses on the local detailed features, such as specific teaching objects or students' and teachers' facial expressions and actions. This effective combination of global and local information can greatly strengthen the model's ability to recognize and classify key teaching elements in a scene of a VR classroom, thereby ensuring high-quality analysis even if the scene is complicated, or the visual signal is unstable.

Figure 4 shows a schematic diagram of the network structure for optimizing VR classroom scene classification.

Specifically in the classification optimization model proposed in this paper, the complex VR classroom scene image is first divided into a series of small blocks and encoded using *Vision Transformer*'s image block embedding idea, a step that provides a basis for understanding the local instructional elements and interaction details by capturing the local features within each block. Next, the *Transformer* encoder receives these encoded image blocks and extracts global features of the entire scene through a self-attention mechanism, which helps the model to capture contextual information about the scene layout and global interactions. At the same time, a branch of image blocks for small-scale features is introduced with the aim of paying special attention to fine-grained elements that may be missed in the global view, such as distant objects or subtle changes in the scene. The recovery of these small-scale features is crucial for understanding and classifying subtle but important interactive behaviors in the teaching environment. To further integrate and enhance the multi-scale information contained in the VR classroom scenes, the model introduces a *Transformer* decoder module. This module enhances the expression strength of multi-scale information in feature encoding through a cross-scale feature fusion strategy. This enables the model not just to accumulate information from different scales but to merge global and local features in a more coordinated and complementary manner, enhancing the overall understanding of VR classroom scenes.

Assuming the image block size of the large-scale branch is represented by o^{LA}, and the small-scale branch by o^{SM}, with the side length of both image blocks denoted by O, and the input image represented by $A \in R^{g*q*3}$. The model initially processes the input VR classroom scene image using convolutional layers of a Convolutional Neural Network (*CNN*). The convolution operation helps extract primary features from the image and divides the image into multiple small blocks through a sliding window approach. This division method preserves local information, laying the foundation for subsequent image block mapping and feature extraction. Further assuming the convolution layer is represented by $J \in R^{f*O*O*3}$ with f convolution kernels of size $O*O$. The task vector is represented by $a_{TA} \in R^{(V+1)*f}$, and the positional embedding vector by $R_{PO} \in R^{(V+1)*f}$. The mapping vectors for the image blocks and a_{TA} are concatenated and added element-wise with R_{PO}, with the following formula representing the calculation of the resultant vector c_0:

$$c_0 = \left[a_{TA}; Q * A\right] + R_{PO} \tag{11}$$

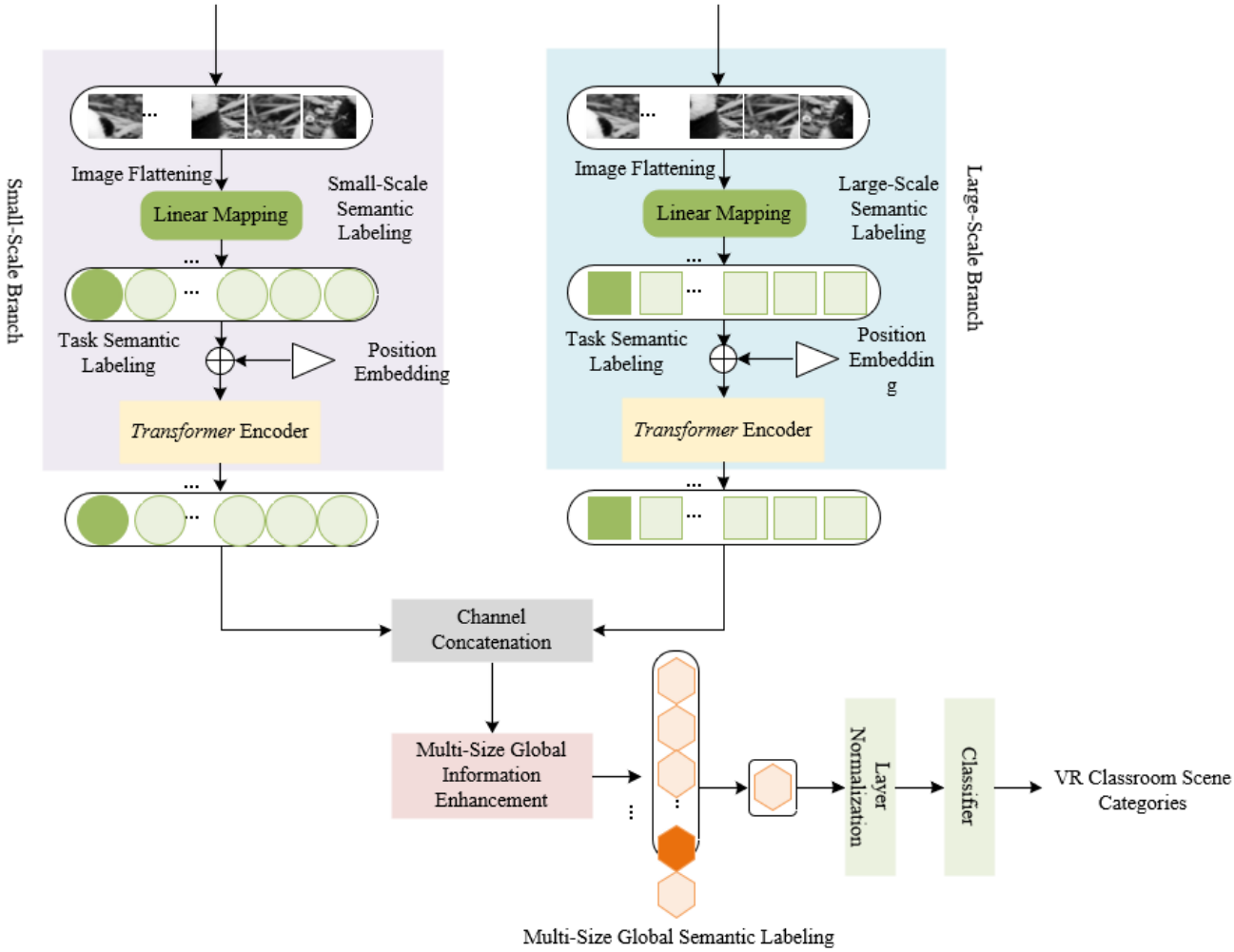

Figure 4. Schematic diagram of network structure for optimization of VR classroom scene classification

Next, each image block obtained through the convolution layer is flattened into a one-dimensional vector. The flattening operation converts the two-dimensional image data into a one-dimensional sequence input for processing by the *Transformer* model. The one-dimensional vectors are then linearly mapped, using a fully connected layer to increase the feature dimensions, better adapting to the self-attention mechanism of the subsequent *Transformer* encoder. The linearly mapped vector sequences are fed into the encoder of the *Transformer* network. Inside the encoder, the features first pass through a multi-head self-attention mechanism module, which allows the model to consider the features of all other image blocks while processing each image block's features, thereby capturing global dependencies. The self-attention mechanism, by computing the relationships among keys, queries, and values, enables the model to dynamically focus on the most relevant parts of the image. The features are then sent to a multilayer perceptron, a fully connected network structure, to further enhance the model's non-linear expression ability and refine and abstract the feature representation. Assuming the output of the encoder on each branch is represented by C_F, and the order of the encoder in the encoder sequence by F, the process expressions are as follows:

$$c_{u'} = MSA\left(LN\left(c_{u-1}\right)\right) + c_{u-1}\left(u = 1, \cdots, F\right) \tag{12}$$

$$c_u = MLP\left(LN\left(c'_u\right)\right) + c'_u\left(u = 1, \cdots, F\right) \tag{13}$$

If the input sequence is represented by $U \in R^{I*F}$, and the learnable matrix by I_{WJN}, the result after the self-attention module is given by:

$$\left[W, J, N\right] = UI_{WJN} I_{WJN} \in E^{F \times 3F_g} \tag{14}$$

$$TX\left(U\right) = SOFTMAX\left(WJ^S \sqrt{F_g}\right) N \tag{15}$$

Assuming the number of "heads" in the multi-head self-attention mechanism is denoted by g, and the learnable matrix by $I_{MSA} \in E^{g*Fg*F}$, the result after the multi-head self-attention mechanism is given by:

$$MSA(U) = \left[TX_1\left(U\right); TX_2\left(U\right); \cdots; TX_g\left(U\right)\right] I_{MSA} \tag{16}$$

The *Transformer* multi-scale global information enhancement module introduced in this paper comprises three parts: an encoding network, a decoding network, and a classifier, as seen in Figure 5. The encoding network, being the core of the *Transformer* structure, is responsible for extracting and encoding the features of input data. Assume the semantic tags for two differently scaled modelings are represented by $c^{LA}{}_F$ and $c^{SM}{}_F$, and the output encoded semantic tags by TO_g. Layer normalization, usually performed before each sublayer of every *Transformer* encoder in the module, aids in stabilizing the training process and accelerating model

convergence. The normalized results are then fed into the multi-head self-attention mechanism module, which processes data in parallel through multiple heads, each learning different representations of the data, thus capturing information from various subspaces. Each sublayer employs residual connections, understood as a form of shortcut operation, allowing the deep network's information to be directly transmitted to subsequent layers, helping to prevent gradient vanishing issues. The output g after the residual connection can be obtained through the following formulas:

$$g' = MSA\left(LN\left(c_F^{LA}; c_F^{SM}\right)\right) + \left[c_F^{LA}; c_F^{SM}\right] \tag{17}$$

$$g = MLP\left(LN\left(g'\right)\right) + g' \tag{18}$$

After the multi-head self-attention module, the multilayer perceptron performs further nonlinear transformations of the features, refining their representation. The final output calculation formula for the encoding network is given by:

$$b = g + \left[c_F^m AR; c_F^t MA\right] \tag{19}$$

The decoding network is similar to the encoder, but the module in this paper includes an additional self-attention layer to combine the outputs of the encoder. In the context of this paper, the decoding network's role is to integrate feature representations of different scales. It enhances the model's ability to express global information by processing features output from the encoding network and combining feature information from different scales.

Finally, after the features are encoded and enhanced through the decoding network, they are passed to the classifier. The classifier, comprising one or more fully connected layers topped with a *Softmax* activation function, maps the features to a probability distribution over predicted categories. In the multi-scale *Transformer* model, the classifier is tasked with synthesizing the enhanced multi-scale global information to achieve precise classification of VR classroom scenes. The final predictions for VR classroom scene categories can be obtained through the following formulas:

$$p = soft\max(FC(LN(token_{task}^{large}; token_{task}^{small}))) \tag{20}$$

$$o = SOFTMAX\left(FC\left(LN\left(TO_{TA}^{LA}; TA_{TA}^{SM}\right)\right)\right) \tag{21}$$

Optimizing the VR classroom scenes based on the category prediction results involves a multi-step, iterative process. Initially, by analyzing the classification predictions and collecting user feedback, areas for improvement are identified. Subsequently, data augmentation and model adjustment strategies are implemented to enhance classification accuracy and model generalization. Following this, the virtual scene's element layout and interaction design are optimized based on the analysis results, enhancing user experience and teaching interaction effectiveness. The new model is then retrained and validated to ensure the effectiveness of the improvements. Lastly, the optimized model is deployed and a continuous monitoring and feedback loop established to ensure ongoing optimization of the VR classroom scenes, better serving teaching and learning activities. This systematic optimization approach ensures continuous improvement of the VR classroom environment, fostering more efficient and engaging teaching experiences.

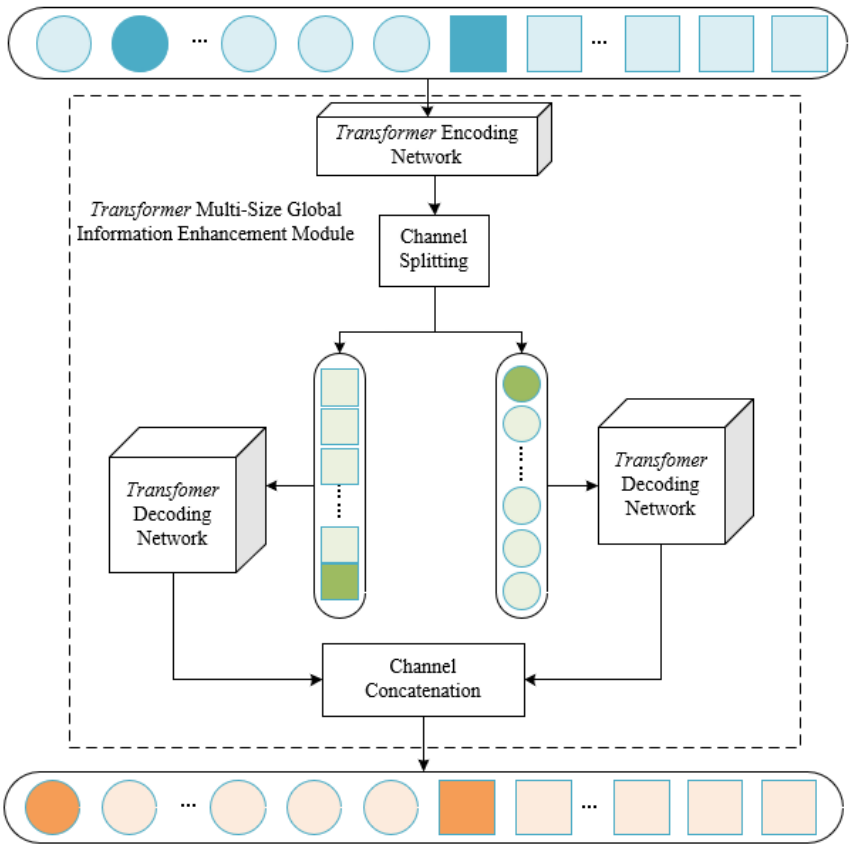

Figure 5. *Transformer* multi-scale global information enhancement module

4. EXPERIMENTAL RESULTS AND ANALYSIS

From Table 1, it is evident that the ablation study results of the VR classroom scene segmentation model demonstrate the performance comparison under different model configurations. *MIoU*, an important metric, is used to measure the accuracy of model segmentation; a higher value indicates better segmentation. *GFLOPs* and *Parameters* represent the computational complexity and size of the model, respectively. As shown, the baseline model serves as the control group, with an *MIoU* of 71.2%, a parameter size of 31.254MB, and computational complexity of 95.26*GFLOPs*. This was the starting point for subsequent improvement. The addition of Multi-Pooling Compression Module 1 to the baseline model resulted in a slight increase in *MIoU* to 71.5% and a slight increase in parameter number and computation complexity, and this indicates that this module has some positive impact on model performance, but the improvement is limited. After adding the multi-pooling compression excitation module 2, the *MIoU* instead decreases to 70.5%, and the amount of computation decreases slightly, this means that module 2 is not as good as module 1 in terms of performance improvement, and even introduces additional disturbances. Later, continuing to add the multi-pooling compression excitation module 3, the *MIoU* improves to 71.7%, which exceeds the performance of the baseline model and the first two modules, while the increase in the number of parameters and the amount of computation is very small, and this indicates that module 3 can enhance model performance while keeping a low computational cost. The introduction of the feature de-hazing branching structure further improves the *MIoU* to 71.6% with a slight increase in the number of parameters but a decrease in the computational effort, and this indicates that the de-hazing branching structure can enhance the model's ability to handle scene details while having little impact on computational efficiency. Finally, incorporating the Feature Distillation Module significantly raised the *MIoU* to 73.1%, the highest among all configurations. Although there was an increase in parameters and computational complexity, the magnitude of performance improvement justifies this increase. In summary, the model proposed in this paper effectively improved the accuracy of VR classroom scene segmentation by introducing multi-pooling compression modules and feature dehazing branch structures, and further significantly enhanced segmentation performance through the feature distillation module. These results validate the effectiveness of the methods presented in this paper, particularly the key role of the feature distillation module in enhancing performance. Even with a slight increase in parameters and computational costs, more accurate segmentation results were achieved.

Table 2 shows the segmentation performance comparison of different models in two different VR classroom scenes. These data are represented in the form of *MIoU* percentage, a commonly used metric for evaluating image segmentation quality. Higher *MIoU* values typically indicate more precise segmentation results. In Scene 1, *HR-Net*'s *MIoU* is 28.9%, suggesting poor performance or lack of assessment in Scene 2. *FDA* performs at 22.4% in Scene 1 and improves in Scene 2, reaching 38.9%, indicating better suitability or effectiveness of the *FDA* model in Scene 2. For *DANN*, the *MIoU* values are 45.3% and 61.2% in the two scenes, respectively, performing better in Scene 2 but not as well as *MF-Net* and *RTF-Net* in Scene 1. *MF-Net* shows good performance in both scenes with *MIoU* values of 48.7% and 64.5%, being the first model to perform well in both scenes. *RTF-Net*'s *MIoU* in Scene 1 is 48.9%, slightly higher than *MF-Net*, but it drops to 58.6% in Scene 2, lower than both *DANN* and *MF-Net*. *BMFF-Net* has a slightly lower performance in Scene 1 at 44.6%, but significantly improves in Scene 2 to 71.2%, surpassing all other models. In Scene 1, this paper's model has an *MIoU* of 48.7%, the same as *MF-Net*, while in Scene 2, it leads all other models with an *MIoU* of 72.1%. It can be concluded that this paper's model has the highest segmentation accuracy in Scene 2 (72.1% *MIoU*), highlighting its robustness and efficiency in complex scenarios. Although not the absolute leader in Scene 1 (matching *MF-Net*), considering the significant performance improvement in Scene 2, it can be inferred that this paper's model has good versatility and adaptability.

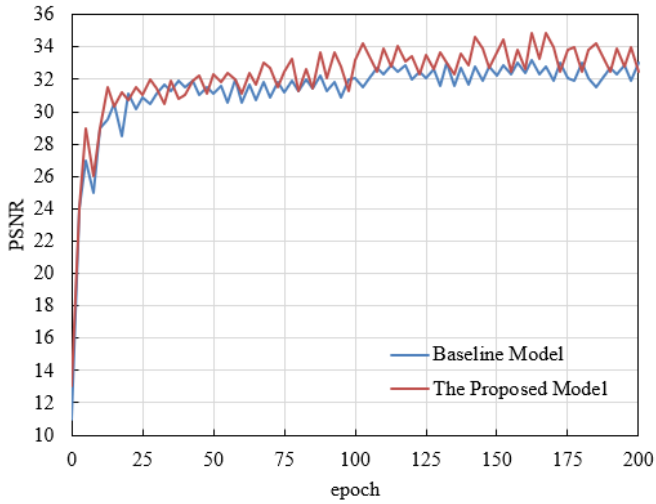

Figure 6. PSNR convergence of the baseline model and the proposed model during training

Figure 6 depicts the Peak Signal-to-Noise Ratio (*PSNR)* values of the baseline model and the proposed model at every 25 epochs during the training process. *PSNR* is a commonly used quality evaluation metric in image processing, often employed to assess the quality of image restoration or image compression. A higher *PSNR* value indicates less distortion and better image quality. From the figure, it is evident that at the initial stage, this paper's model starts with a *PSNR* of 13, while the baseline model is at 11, demonstrating superior performance of the proposed model from the outset. During the mid-phase of training, both the proposed model and the baseline model exhibit a steady increase in *PSNR*, but the proposed model consistently maintains a lead over the baseline model. For example, at epoch 100, the proposed model reaches 32.2, while the baseline model is at 31.1. In the later stages of training, the *PSNR* values of both models tend to stabilize. The proposed model reaches 34.2 at epoch 200, compared to 33 for the baseline model. Not only does the proposed model maintain its leading position, but the margin of lead also increases in the later stages of training. Observing the overall trend, the proposed model shows a more stable and continuous upward trend in *PSNR* growth, while the baseline model experiences slight fluctuations at certain periods, such as a decrease followed by an increase in *PSNR* between epochs 150 to 175. The following conclusions can be drawn: throughout the training process, the proposed model demonstrates superior performance, with not only a higher starting point for *PSNR* but also a more stable growth trend and faster convergence rate. In the later stages of training, the proposed

model achieves higher *PSNR* values than the baseline model, indicating significant improvements in image quality, especially in the later phases of training. Therefore, it can be emphasized that the proposed model is effective and superior in image processing tasks, particularly suitable for applications requiring high *PSNR*.

Table 3 shows the ablation study results of the VR classroom scene classification optimization model, considering the impact of spatial scale information extraction and multi-scale global information enhancement techniques. Accuracy (*Acc*) is a standard metric for evaluating the performance of classification models, and the table presents the accuracies for two different training ratios (20% and 50%). As can be known from the data given in the table, without spatial scale information extraction and multi-scale global information enhancement, the model achieved an accuracy of 94.12% at 20% training ratio, and 95.68% accuracy at 50% training ratio, acting as a baseline of performance comparison. In case that spatial scale information extraction is adopted, there is an improvement in accuracy for both training ratios (reaching 95.63% and 96.12% respectively), and this indicates that even by solely extracting spatial scale information, the model's classification performance has been improved indeed. With the simultaneous application of both spatial scale information extraction and multi-scale global information enhancement, the model's performance further increases, reaching accuracy of 96.35% and 96.98%. This demonstrates the effectiveness of combining these two techniques in enhancing the model's classification capability. The conclusion is that by integrating the *Transformer* architecture, the model effectively extracts multi-scale spatial information and deeply integrates this information through global information enhancement techniques, thereby significantly improving the classification accuracy of VR classroom scenes. The ablation study results indicate that the introduction of each technology positively contributes to model performance, and the combination of both technologies yields the best performance.

Table 4 provides the classification results of different VR classroom scene classification optimization models in various scenes and training ratios, represented by accuracy (%). Each column corresponds to a different scene and training set ratio, and each row represents a specific classification model. From the table, it is clear that the *SSGA-E* model performs best in Scene 2 with an 80% training ratio, achieving an accuracy of 97.87%. *VGG*11 offers data across all scenes and training ratios, showing relatively stable performance, but it's not the highest in most cases. *ViT* reaches an accuracy of 97.25% in Scene 2 with an 80% training ratio, also a strong contender. SA-Gate provides data in all scenes but generally falls short of the proposed model in accuracy. *D-CNN*s reach 96.32% accuracy in Scene 2 with an 80% training ratio but underperform compared to the proposed model in other scenes and ratios. *SF-CNN* excels in Scene 2 with an 80% training ratio, matching *SSGA-E* at 97.87% accuracy, but still falls behind this paper's model in other scenarios and ratios. *ViT-21k* shows impressive performance in Scene 1 with a 50% training ratio at 96.24% accuracy but lacks data for Scene 2 and Scene 3 at a 50% training ratio. The proposed model provides data across all scenes and training ratios and achieves the highest accuracy in most cases. Particularly in Scene 2 with an 80% training ratio, it reaches an accuracy of 98.88%, the highest among all models. It can be concluded that the proposed model demonstrates competitive or optimal performance in all scenes and training ratios. Especially in Scene 2 with an 80% training ratio, achieving an accuracy of 98.88% is the highest among all models. Additionally, its performance is consistently stable in other scenes and ratios, almost always maintaining top-level performance. This indicates that the proposed model has strong generalization capabilities and efficient classification performance.

Table 5 provides performance data of different VR classroom scene classification optimization models in various scenes, assessed through Accuracy (*Acc*) and Intersection over Union (*IoU*) metrics. The table reveals that in the "Historical Recreation" scene, the *VGG*11 model showed the highest accuracy (99.5%) and *IoU* (97%). The proposed model also had high accuracy (98.7%) and *IoU* (96.8%) in this scene, nearly matching the best performance. In the "Geographical Exploration" scene, *VGG*11 again exhibited high accuracy and *IoU*, but proposed model model slightly underperformed with an accuracy of 89.3% and *IoU* of 83.6%. In the "Space Simulation" scene, the *ViT* model stood out with 85.6% accuracy and 65.6% *IoU*.

Table 1. Ablation study results of the VR classroom scene segmentation model

Serial Number	Model	*MIoU*/%	*Parameters*(*MB*)	*GFLOPs*
1	Baseline Model	71.2	31.254	95.26
2	+ Multi-Pooling Compression Module 1	71.5	31.268	95.78
3	+ Multi-Pooling Compression Module 2	70.5	31.257	95.13
4	+ Multi-Pooling Compression Module 3	71.7	31.258	95.47
5	+ Feature Dehazing Branch Structure	71.6	31.698	95.33
6	+ Feature Distillation Module	73.1	31.478	96.86

Table 2. Performance data comparison of different VR classroom scene segmentation models

Serial Number	Model	Scene 1(*MIoU*/%)	Scene 2(*MIoU*/%)
1	*HR-Net*	28.9	-
2	*FDA*	22.4	38.9
3	*DANN*	45.3	61.2
4	*MF-Net*	48.7	64.5
5	*RTF-Net*	48.9	58.6
6	*BMFF-Net*	44.6	71.2
7	The proposed model	48.7	72.1

Table 3. Ablation study results of the VR classroom scene classification optimization model

Spatial Scale Information Extraction	Multi-Scale Global Information Enhancement	*Acc*(%)	
		20%	50%
		94.12	95.68
✓		95.63	96.12
✓	✓	96.35	96.98

Table 4. Comparative results of different VR classroom scene classification optimization models

Method	Scene 1		Scene 2		Scene 3	
Training Ratio	50%	80%	20%	50%	10%	20%
SSGA-E	-	97.87	92.45	95.68	88.95	91.56
*VGG*11	94.25	96.25	87.89	91.78	87.85	87.47
ViT	-	97.25	91.24	93.65	87.26	91.36
SA-Gate	-	95.21	89.64	89.21	89.64	90.14
D-CNNs	-	96.32	91.25	95.87	88.95	91.23
SF-CNN	-	97.87	92.36	95.13	91.23	92.33
ViT-21*k*	96.24	-	94.25	-	91.47	-
The Proposed model	96.87	98.88	95.87	96.89	92.56	92.58

Table 5. Scene comparison experiment of different VR classroom scene classification optimization models

Model Name	Historical Recreation		Geographical Exploration		Space Simulation		Underwater World		Art Gallery	
	Acc	*IoU*	*Acc*	*IoU*	*Acc*	*IoU*	*Acc*	*IoU*	*Acc*	*IoU*
SSGA-E	98.5	96.3	92.1	83.2	83.2	68.6	70.9	59.4	57.9	40.5
*VGG*11	99.5	97	91.4	85.7	74.5	67.9	71.4	58.8	39.5	16.4
ViT	97.1	95.4	92.6	78.9	85.6	65.6	78.4	51.6	57.9	33.8
SA-Gate	97.5	95.6	85	72.3	81.3	58.6	68.7	52.5	57.8	33.5
D-CNNs	98.2	96.8	78.4	64.5	67	59.3	54.3	43.5	37.2	28.8
SF-CNN	99.2	96.3	91.3	85.6	77.8	66.9	72.4	58.2	58.9	42.5
ViT-21*k*	-	98	-	77.8	-	51.4	-	54.5	-	28.9
The Proposed Model	98.7	96.8	89.3	83.6	71.3	64.5	72.3	58.9	52.3	41.2
Model Name	**Science Laboratory**		**Natural Disaster Simulation**		**Digital Visualization**		**Language Immersion**		***Acc***	***IoU***
	Acc	*IoU*	*Acc*	*IoU*	*Acc*	*IoU*	*Acc*	*IoU*		
SSGA-E	24.5	21.5	17.2	3.3	61	41.2	51.4	41.2	61.2	51.2
*VGG*11	27	23.4	12.3	1.8	51.2	45.6	54.7	48.5	56.7	50.2
ViT	51.2	27.8	7	0.8	56.8	15.6	48.9	43.2	63.2	45.6
SA-Gate	23.6	18.9	0	0	56.3	23.8	51.2	47.8	57.4	44.1
D-CNNs	13.2	9.8	0.1	0	31.2	24.7	31	26.9	45.2	38.9
SF-CNN	31.4	23.5	12.3	3.5	41.7	25	72.3	58.6	62.3	51.2
ViT-21*k*	-	24.5	-	14.5	-	38.9	-	44	-	47.5
The Proposed Model	38.5	28.8	55.8	11.2	54.6	45.8	63.8	53.2	65.8	53.2

The proposed model model had lower performance in this scene, with an accuracy of 71.3% and *IoU* of 64.5%. In the "Underwater World" scene, the proposed model performed similarly to *ViT*, with 72.3% accuracy and 58.9% *IoU*. In the "Art Gallery" scene, all models generally showed decreased performance, with *SF-CNN* performing best and the proposed model at a moderate level. For other scenes such as "Science Laboratory," "Natural Disaster Simulation," "Digital Visualization," and "Language Immersion," the proposed model consistently showed relatively high accuracy and *IoU* across all scenes, notably in the "Natural Disaster Simulation" scene, with an accuracy of 55.8% and *IoU* of 11.2%, significantly higher than other models. Overall, while this paper's model was not always the best in individual scenes, it demonstrated robustness and consistency across multiple different scenes, particularly in more challenging scene classification tasks.

5. CONCLUSION

This paper's research primarily focuses on image segmentation and classification issues in VR classroom scenes. By proposing and experimenting with various model optimization methods, the aim is to improve the accuracy and efficiency of scene understanding. The paper first focuses on the segmentation of VR classroom scenes, proposing a method based on feature enhancement and feature distillation, notably introducing multi-pooling compression excitation modules and feature dehazing branch structures to enhance the network's ability to capture effective information and reduce interference from irrelevant information. Secondly, the paper explores classification optimization for VR classroom scenes based on multi-scale global information enhancement, effectively extracting multi-scale spatial information in the scenes and deeply integrating this information through the introduction of the *Transformer* architecture.

Ablation experiments on various components of the model demonstrate the positive impact of introduced multi-pooling compression excitation modules and feature dehazing branch structures on improving segmentation performance. Particularly, the feature distillation module significantly enhances the model's performance at various training stages. Compared with existing models, the proposed model not only performs well in individual scenes but also shows good robustness and generalization across multiple scenes. The

proposed model has been demonstrated to have high accuracy and *IoU* in various VR classroom scenes, especially in more challenging scenes.

The proposed optimization model has been verified to have significant effectiveness in image segmentation and classification tasks. By combining innovative applications of deep learning architectures (such as *Transformer*) and feature enhancement methods, the proposed model can give high accuracy and high *IoU* performance in multiple complex scenarios. Ablation experiment and performance comparison experiment of multiple scenarios can further confirm the robustness and generalization ability of the proposed method. These research attained in this paper results not only provide an effective solution for image processing of VR classroom scenes, but also give a valuable reference for similar research in the field.

ACKNOWLEDGMENT

General Project of Humanities and Social Sciences Research of the Ministry of Education "Research on Intelligent Measurement and Personalized Intervention of Students' Participation in Classroom Based on Video" (Grant No.: 23YJCZH319).

REFERENCES

[1] Liu, Y.T., Cheng, P.Y., Shih, S.P., Huang, T.Y. (2023). MetaClassroom: A WebXR-based hybrid Virtual Reality classroom. 2023 IEEE International Conference on Advanced Learning Technologies (ICALT), Orem, UT, USA, pp. 280-281. https://doi.org/10.1109/ICALT58122.2023.00088

[2] Sharrab, Y., Almutiri, N.T., Tarawneh, M., Alzyoud, F., Al-Ghuwairi, A.R.F., Al-Fraihat, D. (2023). Toward smart and immersive classroom based on AI, VR, and 6G. International Journal of Emerging Technologies in Learning, 18(2): 4-16. https://doi.org/10.3991/ijet.v18i02.35997

[3] Gu, S., Zhang, S., Miao, Y. (2022). Artificial intelligence in construction of English classroom situational teaching mode based on digital twin technology. Wireless Communications and Mobile Computing, 2022: 8357761. https://doi.org/10.1155/2022/8357761

[4] Lai, C., Gao, Q., Zheng, Z., Yuan, D., Zhou, B., Hong, R. (2021). Research on head-up and down behavior computer detection by deep learning and artificial intelligence. In 2021 IEEE 3rd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Changsha, China, pp. 597-600. https://doi.org/10.1109/ICCASIT53235.2021.9633455

[5] Pangsapa, P., Wong, P.P.Y., Wong, G.W.C., Techanamurthy, U., Mohamad, W.S., Shen, J.D. (2023). Enhancing humanities learning with metaverse technology: A study on student engagement and performance. In 2023 11th International Conference on Information and Education Technology (ICIET), Fujisawa, Japan, pp. 251-255. https://doi.org/10.1109/ICIET56899.2023.10111125

[6] Zhu, M., Yang, L., Zhang, Y. (2022). Design and application of project-based teaching of convergence media smart classroom based on VR+AR technology. In 2022 International Conference on Education, Network and Information Technology (ICENIT), Liverpool, United Kingdom, pp. 37-42. https://doi.org/10.1109/ICENIT57306.2022.00016

[7] Li, X., Chen, H., He, S., Chen, X., Dong, S., Yan, P., Fang, B. (2023). Action recognition based on multimode fusion for VR online platform. Virtual Reality, 27(3): 1797-1812. https://doi.org/10.1007/s10055-023-00773-4

[8] Rong, J. (2022). Innovative research on intelligent classroom teaching mode in the "5G" era. Mobile Information Systems, 2022: 9297314. https://doi.org/10.1155/2022/9297314

[9] Zhao, Z., Wu, W. (2022). The Effect of Virtual Reality Technology in Cross-Cultural Teaching and Training of Drones. In: Rau, PL.P. (eds) Cross-Cultural Design. Applications in Learning, Arts, Cultural Heritage, Creative Industries, and Virtual Reality. HCII 2022. Lecture Notes in Computer Science, Springer, Cham, 13312: 137-147. https://doi.org/10.1007/978-3-031-06047-2_10

[10] Nasruddin, Z.A., Mohd Ariffin, N.H., Abdul Rashid, N.S., Mazlin, I. (2023). MYbody: Augmented reality mobile app for understanding body boundaries using MADLC. In 2023 10th International Conference on Electrical and Electronics Engineering (ICEEE), Istanbul, Turkiye, pp. 200-206. https://doi.org/10.1109/ICEEE59925.2023.00044

[11] Amara, K., Zenati, N., Djekoune, O., Anane, M., Aissaoui, I.K., Rahma Bedla, H. (2021). I-DERASSA: E-learning platform based on augmented and Virtual Reality interaction for education and training. In 2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy (AI-CSP), El Oued, Algeria, pp. 1-9. https://doi.org/10.1109/AI-CSP52968.2021.9671151

[12] Wang, B. (2017). Evaluation of sports visualization based on wearable devices. International Journal of Emerging Technologies in Learning, 12(12): 119-126.

[13] Ou Yang, F.C. (2019). The design of AR-based virtual educational robotics learning system. In 2019 8th International Congress on Advanced Applied Informatics (IIAI-AAI), Toyama, Japan, pp. 1055-1056. https://doi.org/10.1109/IIAI-AAI.2019.00224

[14] Gao, C., Wu, Q. (2017). Design and practice of surveying experiment system based on a virtual platform. International Journal of Emerging Technologies in Learning, 12(4): 53-61. https://doi.org/10.3991/ijet.v12i04.6924

[15] Kojima, S., Suetake, N. (2023). Fast and effective scene segmentation method for luminance adjustment of multi-exposure image. IEEE Access, 11: 1128-1140. https://doi.org/10.1109/ACCESS.2022.3233546

[16] Yu, R., Han, L., Zhang, W. (2022). Automatic scene segmentation algorithm for image color restoration. In Proceedings of the 2022 6th International Conference on Electronic Information Technology and Computer Engineering, EITCE 2022, pp. 746-751. https://doi.org/10.1145/3573428.3573777

[17] Su, S., Xu, G., Wang, B. (2023). Research on semantic segmentation of night infrared image for road scene. In Third International Conference on Signal Image Processing and Communication (ICSIPC 2023), Kunming, China, pp. 314-320. https://doi.org/10.1117/12.3004941

[18] Zhao, S., Huang, W., Yang, M., Liu, W. (2023). Real rainy scene analysis: A dual-module benchmark for image deraining and segmentation. In 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Brisbane, Australia, pp. 69-74. https://doi.org/10.1109/ICMEW59549.2023.00018

[19] Ma, Y., Pei, J., Zhang, X., Huo, W., Zhang, Y., Huang, Y., Yang, J. (2023). An optical image-aided approach for zero-shot SAR image scene classification. In 2023 IEEE Radar Conference (RadarConf23), San Antonio, TX, USA, pp. 1-6. https://doi.org/10.1109/RadarConf2351548.2023.10149719

[20] Parseh, M.J., Rahmanimanesh, M., Keshavarzi, P., Azimifar, Z. (2023). Semantic embedding: Scene image classification using scene-specific objects. Multimedia Systems, 29(2): 669-691. https://doi.org/10.1007/s00530-022-01010-9

[21] Qiao, Y., Ge, J., Zhang, Y., Ling, Y. (2023). Remote sensing image scene classification based on transfer learning and swin transformer model. In International Conference on Remote Sensing, Mapping, and Geographic Systems (RSMG 2023), Kaifeng, China, pp. 186-196. https://doi.org/10.1117/12.3010458

[22] Li, Z., Zheng, K., Ni, L., Gao, L. (2023). Level merging attention based dense network for remote sensing image scene classification. In International Conference on Remote Sensing, Mapping, and Geographic Systems (RSMG 2023), Kaifeng, China, pp. 88-93. https://doi.org/10.1117/12.3010658