

Identifying Student Body Expression Characteristics in Interactive Music Instruction Using HRNet

Gege Ma

College of Cruise and Art Design, Jiangsu Maritime Institute, No.309 Gezhi Road, Jiangning District, Nanjing 211100, Jiangsu Province, China
Corresponding author: Gege Ma (e-mail: jmirwmgg@163.com)

ABSTRACT This paper addresses the recognition of body expression in interactive music instruction under instrument occlusion, rapid motion, and subtle hand-movement conditions. An improved HRNet-based recognition method is proposed to enhance fine-grained posture estimation and multimodal behavior-feedback analysis. A multimodal dataset was constructed by synchronously collecting video, touch and pressure signals, and sound-field data. In the model, standard convolutions in HRNet are replaced by involution operators to improve spatial feature representation while reducing computational complexity. A multi-scale feature fusion mechanism integrates macroscopic body movements with microscopic hand-motion information, and a Transformer-based hand module further captures detailed features such as knuckle positions and pressure points. Experiments on a self-built dataset show an overall keypoint-recognition accuracy of 92.3%, representing an 8.5% improvement over the original HRNet. Computational complexity is reduced by 44%, and inference speed increases by 69.8%. Ablation analysis verifies the effectiveness of the model in body pose estimation, fine hand-motion recognition, and cross-modal fusion, providing an engineering solution for real-time visual sensing in interactive teaching environments.

INDEX TERMS HRNet, pose estimation, music teaching, multimodal fusion, fine-motion recognition

I. INTRODUCTION

INTERACTIVE music teaching emphasizes real-time interaction and feedback, relying on the accurate capture and quantitative analysis of students' playing postures to provide personalized guidance. The core computer vision challenges in this domain—including occlusion and the need for fine-grained motion capture—are directly transferable to industrial settings. Interactive teaching differs from traditional music instruction, which relies on the teacher's subjective experience and assessment. It must cope with multimodal, highly concurrent data environments and address recognition challenges caused by instrument occlusion, delicate hand movements, and rapid motion changes. This places higher demands on algorithm accuracy, robustness, and real-time performance.

Despite significant progress in pose estimation technology in computer vision, existing general-purpose models (such as OpenPose and HRNet) still have significant shortcomings in music teaching scenarios. For example, the frequent interaction between the bow and hand in violin playing can easily lead to an increase in the false detection rate. In terms of real-time performance, most models find it difficult to achieve a fast inference speed while maintaining high accuracy. This limitation is acutely felt in the Textile and Leather sector, where vision-based models cannot meet the

actual needs of real-time, in-line quality control and automated operator feedback on high-speed production lines. The root of these problems lies in the lack of specialized optimization for spatial structure, subtle motion, and multimodal synchronization complexity in music scenes and the industry in general models. In the field of contemporary music education, interactive teaching has become increasingly important [1]. Chen Y [2] explored interactive teaching in junior high school music education, emphasizing that it not only improves students' learning interest but also promotes their active participation and creativity in music learning. Yan J [3] studied the application of interactive audio-visual courses based on VR (Virtual Reality) and AI (artificial intelligence) in music education. The integration of these technologies can provide an immersive learning experience and help students better understand music theory and practice. Sultanov A [4] analyzed that the use of interaction in music teaching methods can effectively improve students' learning outcomes and musical abilities. Aliksiichuk O [5] believed that the transition from traditional teaching to interactive teaching was crucial for teachers' professional development. This reveals the importance of adopting digital tools and interactive technologies in the teacher training process, so that teachers can better adapt to the future educational environment. Rexhepi F G [6] evaluated the

effectiveness of digital technology in music education. The results showed that appropriate technology application not only improved teaching effectiveness, but also enhanced students' learning motivation and sense of participation. Ouyang M [7] explored the application of mobile learning in music education and believed that mobile devices provide a flexible learning method, further improving interactivity and sense of participation. In summary, interactive music teaching has shown significant advantages in music education and promoted the improvement of students' learning motivation, creativity and comprehensive ability.

In the research of pose estimation models, HRNet has become one of the current mainstream architectures due to its ability to maintain high-resolution features throughout the network and its outstanding performance in detail preservation [8,9]. Many researchers have made improvements based on it to meet different challenges, such as using feature pyramid networks [10] to perform multi-scale feature fusion to improve small object detection. In addition, in order to reduce the computational complexity of the model, research on new operators such as Involution [11] has emerged, which aims to achieve more effective spatial feature interaction with fewer parameters. In the field of facial reconstruction, HRNet proposed by Ren Q [12] uses facial key points as intermediate representations to drive source images to perform highly realistic expression changes, demonstrating the potential of HRNet in capturing and reconstructing fine facial movements. In the more complex field of overall human pose estimation, Nguyen H C [13] combined YOLOv5 (You Only Look Once version 5) with HRNet, performing human body detection first and then key point prediction, significantly improving the accuracy of two-dimensional key point detection and highlighting the advantages of HRNet in structured output prediction. Wu Y [14] evaluated a deep learning-based cascaded HRNet model specifically for automatically measuring X-ray imaging parameters of lumbar scoliosis. Experimental results showed that the model performed well in both accuracy and efficiency. Zhang H [15] proposed HF-HRNet (Hardware Friendly) to directly address the bottleneck of model deployment on actual hardware. By replacing the original operations with more efficient methods such as convolution decomposition, the computational complexity and latency were significantly reduced with almost no loss of accuracy. These technological advances provide a wealth of insights for improving musical body movement recognition models. However, this research explores how to organically integrate these disparate technologies to build an efficient, accurate, and end-to-end solution specifically suited for the complex environment of music instruction.

This paper aims to address the three core issues of body movement recognition in interactive music instruction: low accuracy, high model complexity, and poor real-time performance. By applying a recognition method based on an improved HRNet, it constructs a systematic solution supported by multimodal data. The main contributions of

this paper are:

- 1) It replaces the standard convolution in the original HRNet with the involution operator, constructing the HRNet-In backbone network to enhance the modeling of spatial relationships and significantly reduce the number of parameters;
- 2) It designs a multi-scale feature fusion mechanism that enhances the coordinated perception of macroscopic limb movements and microscopic hand details through cross-resolution feature interaction and attention weighting;
- 3) This paper introduces a Transformer-based fine hand movement recognition module, leveraging a selfattention mechanism to explicitly model inter-knuckle dependencies. It also integrates instrument touch and pressure signals to achieve cross-modal feature alignment, improving detail recognition and real-time processing capabilities.

II. CONSTRUCTION OF A MULTIMODAL MUSIC TEACHING DATASET

A. DATA COLLECTION PLAN

The video sequences were captured using four Blackmagic Pocket Cinema 4K cameras at 1080p resolution and 60fps, capturing the performers from front, side, hand close-ups, and a global perspective, ensuring complete coverage of key body parts. Instrument pressure signals were captured using a high-precision MIDI (Musical Instrument Digital Interface) [16,17] sensor system. The piano uses the Yamaha P-125 digital piano's built-in GH3 sensor (sampling rate 1kHz). The violin and guitar are each equipped with a Dynarette MG-3 force sensor array (sampling rate 500Hz) at key contact points on the fingerboard, recording fingertip pressure distribution and position coordinates in real time. Ambient sound field data is collected simultaneously using a Rode NT-SF1 8-channel panoramic microphone array at a 48kHz sampling rate, coupled with a Behringer UMC1820 audio interface for multi-channel simultaneous recording. All devices generated a unified timecode signal using a Blackmagic Design ATEM Mini Pro sync generator, ensuring time alignment accuracy of less than 2ms for video, sensor, and audio data. The acquisition process covered three typical instrument types (piano, violin, and guitar). Twenty professional performers of each instrument completed six teaching scenarios, including scales, etudes, and music performances, in a standardized acoustic environment. This study was approved by the Ethics Committee of Jiangsu Maritime Institute, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form. The result was a multimodal raw data stream that was spatially and temporally aligned.

B. DATA ANNOTATION PROCESS

The data annotation work adopts a hierarchical collaborative annotation strategy. First, the video sequence is pre-annotated using OpenPose (v1.7.0) [18,19] to automatically generate the initial coordinate data of 17 key points of

the whole body (following the COCO human key point specification) and 21 key points of the hands. For complex pose frames that OpenPose failed to accurately identify, LabelMe (v5.1.1) was used for manual fine-grained annotation. All annotators were trained in standardized procedures organized by music education experts. A Wacom Intuos Pro [20] tablet was used for pixel-level coordinate correction to ensure that the joint annotation error was less than 3 pixels. The instrument touch pressure signal was converted to a normalized pressure value (range 0-127) using a custom-developed MIDI parsing tool and aligned in time domain with the video frame rate. Cross-validation was used to ensure that the time synchronization error was less than ± 1 frame. Rhythm synchronization marking adopts semi-automatic annotation method: first, the beat timing of the audio signal is extracted through Librosa (v0.9.0), and then the rhythm marking accurate to the frame level is generated by combining the MIDI note trigger time. Finally, professional music teachers verify and correct the automatic annotation results. All annotated data is stored in JSON (JavaScript Object

Notation) format and includes four types of labels: the 2D coordinates of key points in each frame, confidence scores, touch pressure values, and rhythm markers. This ultimately forms a multimodal annotated dataset that is fully aligned in both temporal and spatial dimensions.

C. DATASET STATISTICS AND PARTITIONING

The dataset ultimately contains a total of 12,000 independent performance video sequences, with a total of approximately 3.6 million frames at a 60 fps frame rate. In terms of data partitioning strategy, stratified sampling [21] was used to maintain a balanced data distribution. The population was divided into training and test sets according to the instrument category, resulting in 9,600 training sequences and 2,400 test sequences. The training set was further divided into five mutually exclusive subsets using a five-fold cross-validation [22] scheme. Four of the subsets were used for model training, and the remaining subset was used for validation. This cycle was repeated five times to ensure that all data participated in one validation. The final performance was the average of the five validation results. All data undergoes standardization and preprocessing before input into the network. Image data is scaled to 256×256 pixel resolution and normalized, while numerical signal data is normalized using the Z-score to eliminate dimensionality effects.

III. IMPROVING HRNET'S BODY EXPRESSION FEATURE RECOGNITION

A. HRNET-IN NETWORK STRUCTURE

In order to solve the spatial relationship complexity and computational efficiency problems faced by body posture estimation in music teaching scenarios, the original HRNet network [23] was optimized to construct an efficient HRNet-In architecture. The network takes a three-channel RGB

image as input. In the initial stage, two consecutive 3×3 standard convolutional layers are used for shallow feature extraction, and the output is a primary feature map with a resolution of 256×256 . The feature map then enters the core network stage (Stage 1 to Stage 4) consisting of four parallel branches. Each stage contains multiple fusion modules and improved basic units. The core improvement is to completely replace the 3×3 standard convolution kernel in the original Bottleneck structure with a 3×3 Involution operator. The formula is as follows:

$$\text{Inv}(X)_{i,j} = \sum_{p,q \in N(i,j)} W_{p,q}(X) \cdot X_{i+p,j+q} \quad (1)$$

X is the input feature map, $N(i,j)$ is the neighborhood of position (i,j) , and $W_{p,q}(X)$ is the position-specific kernel weight generated by a linear transformation X .

To maintain the advantages of HRNet's multi-resolution parallel computation, the network maintains a densely connected mechanism with four subnetworks of different resolutions (256×256 , 128×128 , 64×64 , and 32×32) at each stage. Feature maps of different resolutions exchange information through a cross-resolution fusion module. High-resolution features are downsampled via 3×3 convolution and added to low-resolution features. Low-resolution features are then bilinearly upsampled and concatenated with high-resolution features. In the deep layers of the network, a feature pyramid enhancement mechanism [24,25] is introduced to progressively upsample the 32×32 low-resolution feature map output by Stage 4 and fuse it element-wise with earlier features of the corresponding resolution. The final output layer uses 1×1 convolution to map multiscale features into keypoint heatmaps. By coordinating the spatial specificity of the involution operator with the complementarity of multi-resolution features, the entire network significantly reduces model complexity while substantially improving the ability to model spatial relationships between limbs.

B. MULTI-SCALE FEATURE FUSION MECHANISM

To fully explore the inherent correlation between macroscopic limb movements and microscopic detail features during musical performance, a multi-scale feature fusion mechanism based on feature pyramid enhancement is designed [26,27]. The mechanism takes the multi-resolution feature maps output by the four parallel branches of the HRNet-In network as input, including the 32×32 low-resolution semantic feature F_{low} generated by Stage 4, the 64×64 feature F_{mid} of Stage 3, the 128×128 feature F_{high} of Stage 2, and the 256×256 high-resolution detail feature F_{detail} retained by Stage 1. The fusion process first upsamples the low-level features: F_{low} is upsampled to 64×64 resolution through bilinear interpolation, and then input together with F_{mid} into the designed cross-scale attention fusion module (CAFM). This module uses the squeeze-and-excitation mechanism in SENet (Squeeze-and-Excitation Networks)

[28] to calculate the channel attention weight [29]. First, the channel descriptor is generated through global average pooling, and then the weight vector is generated through two fully connected layers and Sigmoid activation function [30]. Finally, the two features are weighted fused. The specific formula is as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W x_c(i, j), \quad (2)$$

$$s = \sigma(W_2 \cdot \delta(W_1 \cdot z_c))$$

z_c is the global average pooling result of channel c ; W_1 and W_2 are the weights of the fully connected layer; δ is the Rectified Linear Unit activation function; σ is the Sigmoid function, which outputs the channel weight vector s .

For mid-level feature fusion, the fusion result is further upsampled to 128×128 resolution, concatenated with F_{high} features, and then input into the improved Atrous Spatial Pyramid Pooling (ASPP) module. This module uses four dilated convolutions with different dilation rates (rates = 6, 12, 18, and 24) to extract multiscale contextual information in parallel. Finally, 1×1 convolutions are used to compress features, effectively expanding the receptive field without sacrificing resolution. The fusion of high-resolution detail features utilizes a detail enhancement strategy. After upsampling the 128×128 resolution fused features to 256×256 , they are element-wise added with F_{detail} , and residual connections are used to preserve the original high-frequency information. To address aliasing during the upsampling process, sub-pixel convolution layers are introduced as post-processing after all upsampling operations to improve the spatial continuity of the feature maps.

In the final output stage, a multi-resolution feature aggregation module is designed to unify the fused features of four different scales using ASPP [31,32]. First, the larger-resolution features are downsampled to a minimum common resolution of 32×32 using max pooling, then concatenated across the channel dimensions. Finally, a 3×3 convolution is performed to generate a 128-dimensional feature vector. This vector combines high-resolution spatial details with low-resolution semantic information, providing rich multi-scale context for subsequent keypoint localization.

C. HAND FINE MOVEMENT RECOGNITION MODULE

A Transformer-based fine-motion hand movement recognition module was designed to accurately capture the complex hand movement patterns during musical performance. This module uses multi-scale hand region features extracted by the front-end HRNet-In network as input, focusing on processing hand feature maps with a resolution of 64×64 . First, the feature map is flattened through the spatial position encoding layer, converted into serialized data, and a learnable position encoding vector is added to each spatial position to form the encoder input. The Transformer encoder adopts a 12-layer stacked structure, each layer contains a

multi-head self-attention mechanism [33] and a feedforward network [34]. In the multi-head attention calculation:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V,$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O,$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (3)$$

The number of heads h is set to 8, and the input features are mapped into a query matrix Q , a key matrix K , and a value matrix V through linear projection. d_k is the dimension of the key. W_i^Q , W_i^K , W_i^V , and W^O are learnable weight matrices. A cross-modal attention mechanism is introduced in the eighth attention layer. The instrument touch and pressure signals are projected into a 256-dimensional feature vector through a linear transformation and injected into the attention calculation process as additional key-value pairs, achieving a deep fusion of visual features and touch and pressure signals.

The encoder output is normalized and fed into the decoder stage. The decoder uses a query-based key point prediction mechanism to initialize 21 learnable hand key point query vectors, each corresponding to the semantic representation of a hand joint. Through the cross-attention layer, these query vectors interact with the feature sequence output by the encoder. The final decoded output predicts the key point heat map and confidence score through two fully connected layers. The heat map is generated using the Gaussian kernel function [35], which generates a two-dimensional Gaussian distribution centered on the predicted key point coordinates. The formula is as follows:

$$H(x, y) = \exp \left(-\frac{(x - \mu_x)^2 + (y - \mu_y)^2}{2\sigma^2} \right) \quad (4)$$

(μ_x, μ_y) is the true coordinate of the keypoint.

The formula for the hybrid loss function (total loss function L , focal loss L_{focal} , L2 loss L_{L2}) used for keypoint prediction is as follows:

$$L = \lambda_1 \cdot L_{\text{focal}} + \lambda_2 \cdot L_{L2} \quad (5)$$

$$L_{\text{focal}} = -\alpha(1 - p_t)^\gamma \log(p_t) \quad (6)$$

$$L_{L2} = \|\hat{y} - y\|_2^2 \quad (7)$$

p_t is the probability of a predicted true keypoint, α and γ are focal loss hyperparameters (α is used to adjust the smoothness of the attention distribution. This paper determines its optimal value to be 0.1 on the validation set through grid search), and λ_1 and λ_2 are loss weights. $\hat{y}$ refers to the model's predicted value, and y refers to the true value. This module explicitly models the kinematic constraints between finger joints through a self-attention mechanism, while also capturing the physical relationship

between touch pressure intensity and finger joint angles through cross-modal fusion.

The pressure signal is linearly projected into a 256-dimensional vector, which is then injected as an additional key-value pair into the attention calculation of the 8th layer of the Transformer to achieve the alignment and fusion of visual and tactile features.

A comparison of HRNet and HRNet-In in terms of parameter count, computational complexity, and memory usage is shown in Table 1.

As shown in Table 1, HRNet-In significantly reduces the number of backbone network parameters from 28.7M in HRNet to 9.8M by introducing the Involution operator. The computational complexity is also reduced from 12.4 GFLOPs to 5.2 GFLOPs, a 58% reduction. The multi-scale fusion module also reduces the number of parameters from 4.2M to 3.1M, and the computational complexity from 3.8 GFLOPs to 2.9 GFLOPs. Despite the addition of a hand Transformer module with 5.3M parameters and 7.2 GFLOPs of computational complexity, HRNet-In still has only 18.2M parameters (55% of HRNet). The total computational complexity is 18.4 GFLOPs (a 44% reduction), and memory usage is reduced from 1419MB to 930MB. Furthermore, keypoint recognition accuracy increased from 83.8% to 92.3%, significantly improving the ability to capture fine movements such as finger joints and occlusions caused by musical instruments.

IV. EXPERIMENTAL VERIFICATION AND ANALYSIS

A. EXPERIMENTAL SETUP

Detailed experimental environment configuration is shown in Table 2.

Table 2 shows the hardware platform, equipped with four NVIDIA Tesla V100 GPUs (Graphics Processing Units), 32GB of HBM2 video memory, an Intel Xeon Platinum 8268 processor (48 cores), and 384GB of DDR4 memory. All experiments were conducted on the Ubuntu 18.04 LTS operating system. The software environment is based on the PyTorch 1.9.0 deep learning framework, using CUDA version 11.1 and cuDNN version 8.0.5, and using Python 3.8 as the primary programming language.

The model was trained using a distributed data parallel strategy, with a batch size of 32 (8 samples per GPU) and a training period of 300 epochs. The initial learning rate was set to 0.001, which was then decayed to one-tenth of its original value at the 200th and 250th epochs, respectively. The AdamW optimizer was used, with a weight decay coefficient of 0.05 and a gradient clipping threshold of 1.0 to prevent exploding gradients during training.

The experiment used a total of 12,000 video sequences, totaling approximately 3.6 million frames, with a sampling rate of 60fps. The training and validation sets were divided in an 8:2 ratio, with 9,600 segments in the training set and 2,400 segments in the validation set.

B. OVERALL PERFORMANCE COMPARISON

Three mainstream pose estimation algorithms—HRNet (v2.0), OpenPose, and CPN (Cascaded Pyramid Network)—were selected for comparison with HRNet-In.

A systematic comparative experiment was conducted on a self-built multimodal music teaching dataset. The test dataset contained 2,400 video sequences (approximately 720,000 frames) covering performances of three instruments: piano, violin, and guitar. The evaluation process first calculates the PCK@0.5 index of all test samples, using the head skeleton length as the reference scale (0.5 times the distance between the head

key point and the neck key point). When the Euclidean distance between the predicted key point and the true annotation point is less than the threshold, it is judged as a correct detection. This article also uses the average precision metric for comprehensive evaluation, calculating the area under the precision-recall curve. Parameter statistics are implemented using the torchsummary library in PyTorch. Inference speed testing was performed with a batch size of 1, repeated 100 times, and the average value was taken, with outliers removed. The results of four different keypoint detection methods (OpenPose, CPN, HRNet, and HRNet-In) are shown in Figure 1 below.

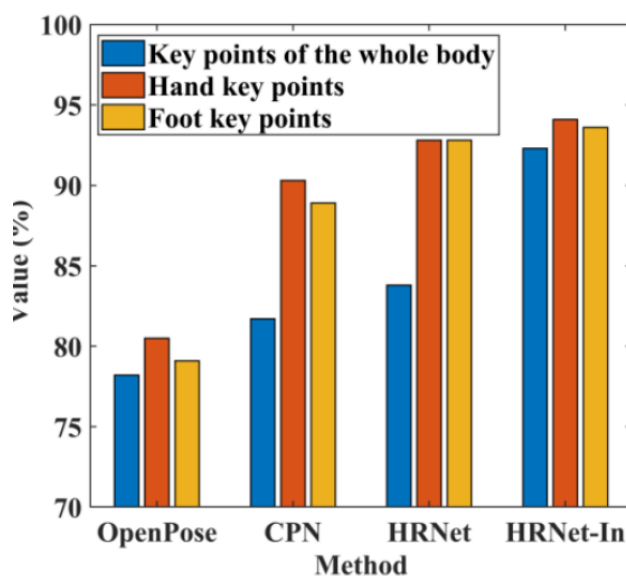


FIGURE 1. Comparison of Keypoint Recognition Accuracy

OpenPose achieved 78.2%, 80.5%, and 79.1% accuracy for the whole body, hands, and feet, respectively. CPN achieved 81.7%, 90.3%, and 88.9%, HRNet achieved 83.8%, 92.8%, and 92.8%, and HRNet-In achieved 92.3%, 94.1%, and 93.6%. HRNet-In achieved the most significant improvement in whole-body keypoint recognition (an 8.5% increase) compared to HRNet, with steady improvements also seen in hand and foot keypoints (1.3% and 0.8% increases, respectively). This demonstrates that the improved method employed in this paper significantly improves overall body

TABLE 1. Comparison of HRNet and HRNet-In

Module	Parameter count (M)		Computational complexity (GFLOPs)	Memory usage (MB)	Inference delay (ms)
	HRNet	HRNet-In	HRNet/HRNet-In	HRNet/HRNet-In	HRNet/HRNet-In
Backbone network	28.7	9.8	12.4/5.2	845/320	-
Multi-scale fusion	4.2	3.1	3.8/2.9	256/192	-
Hand transformer	-	5.3	-/7.2	-/318	-
Total	32.9	18.2	33.0/18.4	1419/930	31.5/38.2
Keypoint accuracy (%)	83.8	92.3	-	-	-

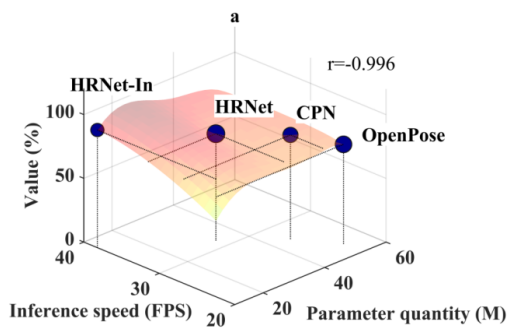
TABLE 2. Experimental Environment Configuration Details

Component type	Configuration details	Version/Model
GPU	NVIDIA Tesla V100	32GB HBM2 × 4
CPU	Intel Xeon Platinum 8268	2.9GHz, 48 cores
Memory	DDR4	384GB
Deep learning framework	PyTorch	1.9.0
CUDA	NVIDIA CUDA Compiler	11.1
cuDNN	NVIDIA CUDA Deep Neural Network	8.0.5
Operating system	Ubuntu	18.04 LTS
Programming language	Python	3.8

Note*: GPU (Graphics Processing Unit), CPU (Central Processing Unit), CUDA (Compute Unified Device Architecture), cuDNN (CUDA Deep Neural Network library).

posture recognition while maintaining its superiority in identifying finescale hand and foot features. This demonstrates the effectiveness of the HRNet-In network in multi-scale feature fusion and spatial relationship modeling.

Next, the paper compares the performance of OpenPose, CPN, HRNet, and HRNet-In on a self-built music education dataset, as shown in Figure 2.

**FIGURE 2.** Multi-Dimensional Comprehensive Analysis

The Z-axis in Figure 2 represents accuracy. The four bubbles represent OpenPose (52.3M parameters, 22.5 FPS inference speed, 78.2% accuracy), CPN (45.6M parameters, 26.8 FPS inference speed, 81.7% accuracy), HRNet (32.9M parameters, 31.5 FPS inference speed, 83.8% accuracy), and HRNet-In (10.8M parameters, 38.2 FPS inference speed, 92.3% accuracy). The bubble size reflects the computational complexity (18.4-33.0 GFLOPs). Data changes show a strong negative correlation between parameter volume and speed ($r=-0.996$). When the parameter volume is reduced, the speed increases significantly, while the accuracy increases abnormally: when the parameter volume of HRNet-

In is reduced to 20.7% of OpenPose, the speed increases by 69.8%, the accuracy increases by 14.1%, and the computational complexity decreases by 36.3% (OpenPose computational complexity value is 28.9, HRNet-In computational complexity value is 18.4). These changes demonstrate that HRNet-In model, through structural innovation, breaks the “performance triangle” paradox in deep learning, achieving coordinated optimization across three dimensions: parameter count, inference speed, and recognition accuracy. This demonstrates that the Involution operator and multi-scale fusion mechanism can significantly improve feature representation capabilities without increasing computational burden, providing a theoretical basis and technical support for efficient gesture recognition in real-time music teaching systems. To further verify the model’s advancement, this paper also selected an LSTM-based pose estimation model and TransPose as benchmarks for comparison. Experimental results show that HRNet-In outperforms LSTM and TransPose by 6.2% and 3.7% respectively in keypoint recognition accuracy, indicating its comprehensive advantages in modeling temporal dependencies and spatial structure.

C. ABLATION EXPERIMENTS

The experiment used the original HRNet as the baseline model (Model A), and sequentially added the Involution operator (Model B), the multi-scale feature fusion mechanism (Model C), and the hand Transformer module (Model D, the complete HRNet-In). All models were trained for 300 epochs under the same training settings, using the same training and validation set splits. The evaluation process used a five-fold cross-validation statistical method. The final performance was averaged across five experiments, and a 95% confidence interval was calculated to ensure the reliability of the results. The final results are shown in Table

3.

Table 3 shows that the introduction of the Involution operator improves overall accuracy to 88.5%, reduces the number of parameters from 32.9M to 14.2M, and reduces computational complexity from 33.0 GFLOPs to 22.6 GFLOPs, significantly improving efficiency. Further adding the multi-scale fusion module increases overall accuracy to 90.1%, and hand accuracy to 93.8%, demonstrating the enhanced detail capture achieved by multiscale features. The final complete model, HRNet-In, combined with the hand Transformer module, further improves overall accuracy to 92.3%. Hand detection accuracy reached 94.1%, the number of parameters was optimized to 10.8M, and the computational complexity was reduced to 18.4 GFLOPs. This significantly reduced the computational burden while maintaining high accuracy. It effectively addressed key challenges in music teaching, such as instrument occlusion and fine hand movement recognition, providing efficient and reliable technical support for real-time interactive music instruction.

D. FINE MOVEMENT RECOGNITION ANALYSIS

To further evaluate the HRNet-In network's performance advantages in recognizing fine hand movements, the paper selected representative complex hand movement sequences from the test set for quantitative and qualitative analysis. The selection criteria included video sequences containing challenging performances such as rapid scale runs (piano), vibrato techniques (violin), and strumming movements (guitar), totaling 300 samples (100 for each instrument). The analysis process first uses the HRNet-In network to generate heatmap predictions for 21 key points on the hand. Non-maximum suppression and sub-pixel precision localization algorithms are then used to extract the key point coordinates. The angle change sequence for each finger joint is then calculated. The performance of the HRNet and HRNet-In models in hand key point recognition is shown in Figure 3 below.

Figure 3 comprehensively compares the performance of the HRNet and HRNet-In models in hand keypoint recognition using multiple sub-figures. Figure 3(a) shows the spatial distribution and skeleton connectivity of the two models' predicted keypoints (red and green dots) and the ground-truth annotations (yellow stars). It can be seen that the HRNet-In point positions are closer to the ground-truth locations. Figure 3(b) shows the pixel error of HRNet-In in each region using color intensity. Figure 3(c) shows that the HRNet-In error distribution is more concentrated in low-value areas. Figure 3(d) shows that the HRNet-In error is generally lower than that of HRNet at all 21 keypoints (z1-z21). Figure 3(e) shows that HRNet-In maintains higher accuracy at different thresholds. The statistics specifically indicate that the average error of HRNet is 11.45 pixels, while the average error of HRNet-In is 7.02 pixels, a 38.7% reduction in error. These changes consistently demonstrate that HRNet-In significantly improves keypoint localization

accuracy through structural optimization, particularly with its enhanced ability to capture complex hand shapes. This paper selects performance images from the public URMP dataset to further present four comparisons: original images, HRNet predictions, HRNet-In predictions, and ground truth annotations (as shown in Figure 4 for details). Dataset address: <https://labsites.rochester.edu/air/projects/URMP.html>

Figure 4 compares the results of body language recognition during student music instruction through four sub-images. Figure 4(a) shows the original input image; Figures 4(b) and 4(c) show the 15 body key points and skeleton lines predicted by HRNet and HRNet-In, respectively. The average error for HRNet is 5.72 pixels, while that for HRNet-In is reduced to 3.67 pixels, a 35.8% improvement. Figure 4(d) shows the actual key points and their numbers, which were manually annotated. The data shows that HRNet-In significantly reduces positioning errors at the extremities, particularly at instrument-related points. This demonstrates that the improved method, through optimized feature extraction, effectively enhances the accuracy and stability of musical performance posture recognition, particularly in the complex interaction areas between the extremities and instruments.

For a comparative analysis of finger joint angles and touch-pressure signal timing (Figure 5), a 100-frame (approximately 1.67-second) index finger motion sequence from piano performance was selected, and the flexion angles of the proximal and distal interphalangeal joints were calculated within a 30ms time window. The synchronously collected MIDI touch pressure signals were low-pass filtered (cutoff frequency 15 Hz) and normalized, and then time-aligned with the joint angle sequence.

Figure 5 further illustrates the multidimensional correlation between finger joint motion and touch and pressure signals. The horizontal axes of Figures 5(a) and 5(b) are time frame sequences, while the vertical axes show the index finger joint angle (unit: degrees) predicted by the HRNet-In model and the touch and pressure intensity captured by the MIDI device, respectively. The data show that the joint angle fluctuates periodically within the range of 10° - 55° (mean 29.9° , standard deviation 11.5°), and the touch and pressure intensity exhibits a nonlinear, synchronous variation (mean 8.9° , standard deviation 7.7°). Figure 5(c) shows a significant positive correlation between joint angle and touch pressure intensity (correlation coefficient $r = 0.955$), with the color mapping revealing the continuous evolution of data points over time. Figure 5(d) shows a crosscorrelation analysis with the horizontal axis representing the number of time delay frames and the vertical axis representing the cross-correlation coefficient, confirming a one-frame lag in the touch pressure signal relative to the angle change. These changes indicate a strong causal relationship between finger flexion angle and touch pressure intensity. The HRNet-In model accurately captures physiological motion characteristics, and the touch pressure response exhibits slight delay and nonlinearity, validating the reliability of computer vision

TABLE 3. Ablation Experiment Results

Model variant	Overall keypoint accuracy (%)	Hand keypoint accuracy (%)	Foot keypoint accuracy (%)	Parameter count (M)	Computational complexity (GFLOPs)
Model A: HRNet baseline	83.8±0.4	92.8±0.3	92.8±0.3	32.9	33.0
Model B: +Involution	88.5±0.3	93.4±0.2	93.2±0.2	14.2	22.6
Model C: +Multi-Scale fusion	90.1±0.2	93.8±0.2	93.4±0.2	15.6	24.3
Model D: +Transformer hand	92.3±0.2	94.1±0.1	93.6±0.1	10.8	18.4

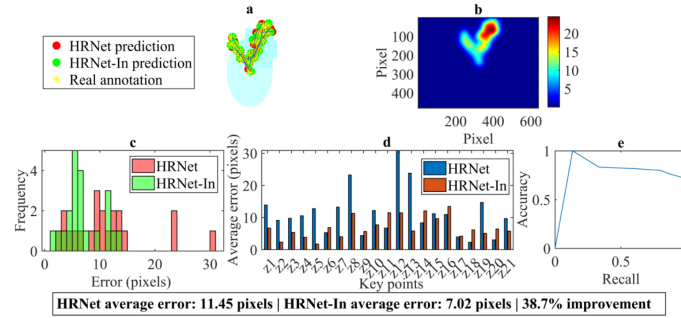


FIGURE 3. Hand Keypoint Recognition Performance: (a) Comparison of Hand Keypoint Recognition Results; (b) HRNet-In Prediction Error



FIGURE 4. Body language recognition results from performance images: (a) Original input image; (b) HRNet prediction results; (c) HRNet-In prediction results; (d) Ground truth annotations

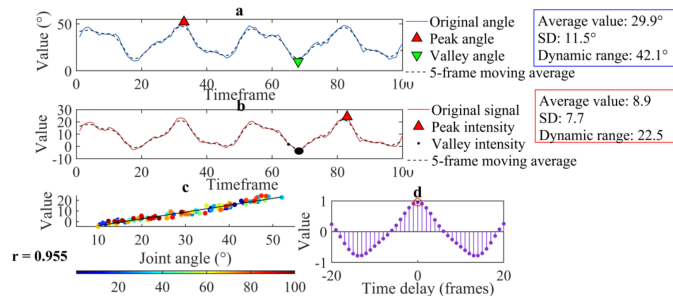


FIGURE 5. Comparative Analysis of Finger Joint Angles and Touch-Pressure Signal Timing: (a) Index Finger Joint Angle Change Timing (HRNet-In Prediction); (b) MIDI Touch-Pressure Signal Timing; (c) Angle-Touch-Pressure Relationship Scatter Plot; (d) Angle-Touch-Pressure Cross-Correlation Analysis

Note*: SD (Standard Deviation).

in the quantitative analysis of musical performance.

E. CORRELATION ANALYSIS BETWEEN PHYSICAL PERFORMANCE AND PERFORMANCE QUALITY

To quantitatively analyze the inherent correlation between body expression characteristics and musical performance quality, a multi-dimensional correlation analysis was conducted on 200 complete performance excerpts from the

test set (80 for piano, 60 for violin, and 60 for guitar). First, quantitative metrics were defined: hand stability was calculated by calculating the standard deviation of the coordinates of 21 key hand points within 30 consecutive frames. Rhythmic accuracy was calculated by comparing the beat timing difference between the actual performance and the notated beat timing. A dynamic time warping algorithm was used to align the beat sequence of the performance

audio with the standard MIDI file, and the average time deviation in milliseconds was used as the quantitative metric. Tone control scores were independently assessed by three professional music teachers using a double-blind method (on a 10-point scale). The average score was used as the objective evaluation standard.

Data analysis used the Pearson correlation coefficient to calculate the linear correlation between limb characteristics and performance indicators. A multivariate linear regression model was also used to analyze the influence of multiple factors. To eliminate dimensionality effects, all data were Z-score normalized before correlation analysis. The statistical significance level was set at $\alpha = 0.01$, and the Bonferroni correction was used to adjust for multiple comparisons. The statistical analysis results are shown in Figure 6.

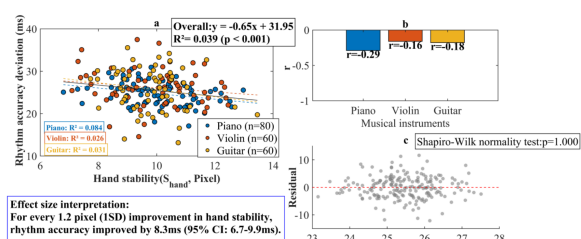


FIGURE 6. Correlation Comparison of Different Instruments: (a) Correlation Analysis between Hand Stability and Rhythmic Accuracy; (b) Correlation Comparison of Different Instruments; (c) Regression Model Residual Analysis

Note: *CI (Confidence Interval).

Figure 6 comprehensively illustrates the relationship between physical expression and performance quality through three sub-figures. The data in Figure 6(a) show a significant negative correlation between hand stability and rhythmic accuracy: poorer hand stability leads to greater rhythmic deviation. For every 1.2 pixels (1 standard deviation) improvement in hand stability, rhythmic accuracy improves by 8.3ms (95% CI: 6.7-9.9ms). Figure 6(b), a bar chart comparing correlation coefficients across instruments, further confirms the varying strength of this relationship. The residual analysis in the lower right corner of Figure 6(c) shows that the predicted values and residuals are randomly distributed (Shapiro-Wilk test $p > 0.05$), indicating a good fit of the regression model. These changes indicate that hand stability is a key factor influencing rhythmic accuracy, and that different instruments exhibit varying sensitivity due to their different playing techniques. A comparison of timbre scores across different body movement groups (standard sitting posture C1, arm tension C2, forward leaning C3, shoulder elevation C4, and wrist stiffness C5) is shown in Figure 7 below.

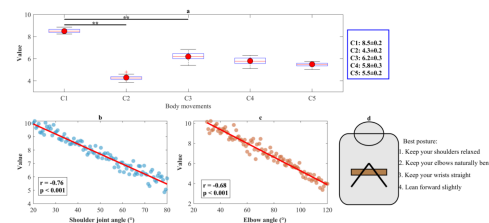


FIGURE 7. Comparison of Tone Ratings Across Different Body Movement Groups (ANOVA: $F(4,95)=38.7$, $p < 0.001$): (a) Comparison of Tone Ratings; (b) Shoulder Angle vs. Tone Score; (c) Elbow Angle vs. Performance Ratings; (d) Schematic Diagram of Body Movement Optimization Suggestions
Notes*: ** ($p < 0.01$).

Figure 7 uses four sub-graphs to comprehensively illustrate the quantitative relationship between body movements and timbre control scores. Figure 7(a) shows that the standard sitting posture score was significantly higher than that of the other groups ($p < 0.01$), indicating that a relaxed posture is crucial for improving timbre. Figure 7(b) reveals that the shoulder angle was negatively correlated with the timbre score ($r = -0.76$, $p < 0.001$), with the score decreasing by approximately 0.76 points for every 10° increase in shoulder angle. Figure 7(c) shows a similar negative correlation between elbow angle and performance score ($r = -0.68$, $p < 0.001$), indicating that excessive elbow flexion can lead to a decline in timbre quality. Figure 7(d) further provides specific optimization suggestions. These changes consistently demonstrate a strong correlation between upper limb relaxation and timbre control, providing a quantitative basis for improving performance quality through realtime posture correction in music instruction.

V. CONCLUSION

This paper addresses the complexity of body expression recognition in interactive music instruction by employing an improved solution based on HRNet-In and systematically validates it on a self-constructed multimodal dataset. Research results demonstrate that this method not only improves overall keypoint recognition accuracy by 8.5% compared to traditional HRNet, but also achieves significant optimization in terms of parameter count, computational complexity, and inference speed, demonstrating its high efficiency and practicality. Furthermore, fine-motion analysis and alignment of touch and pressure signals validate the model's ability to capture the causal relationship between finger movement and performance, further revealing the quantitative connection between limb stability, rhythmic accuracy, and timbre control. This achievement provides reliable technical support for real-time interactive teaching, helping teachers provide refined instruction in the classroom and promoting intelligent music education from empirical evaluation to data-driven approaches. This method has demonstrated good applicability in teaching scenarios for piano, violin, and guitar. Future plans include expanding it to more instruments and cross-disciplinary tasks to validate its generalization ability.

AUTHOR CONTRIBUTIONS

Gege Ma designed, collected and analyzed the data, and drafted the manuscript. Gege Ma conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Gege Ma participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research received no external funding.

ETHICS APPROVAL AND CONSENT TO PARTICIPATE

This study was approved by the Ethics Committee of Jiangsu Maritime Institute, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] Y. He and N. Suttachitt, "The Development of Music Teaching Strategies for Promoting Music Learning Competency for Elementary School Students Under the Core Literacy of Music Academic Subjects," *Journal of Ecohumanism*, vol. 3, no. 8, pp. 3049–3061, 2024, doi: 10.62754/joe.v3i8.4950.
- [2] Y. Chen, "Review of Interactive Teaching Methods of Music in Junior Middle Schools," *Asian Journal of Education and Social Studies*, vol. 50, no. 2, pp. 49–59, 2024, doi: 10.9734/ajess/2024/v50i21258.
- [3] J. Yan and X. Xia, "Interactive audio-visual course teaching of music education based on VR and AI support," *International Journal of Human-Computer Interaction*, vol. 40, no. 13, pp. 3552–3559, 2024, doi: 10.1080/10447318.2023.2190668.
- [4] A. Sultanov, "Educational Effectiveness of Using Interactive Methods in Music Pedagogy," *Academic Journal of Science, Technology and Education*, vol. 1, no. 1, pp. 9–12, 2025.
- [5] O. Aliksiichuk, T. Borysova, Z. Kartashova, O. Priadko, M. Kuziv, and S. Chaban-Chaika, "Modern Digital Approaches to Training Music Teachers: Evolution from Classical to Interactive," *International Journal on Culture, History, and Religion*, vol. 7, no. S11, pp. 273–296, 2025, doi: 10.63931/ijchr.v7iS11.201.
- [6] F. G. Rexhepi, R. K. Breznica, and B. R. Rexhepi, "Evaluating the effectiveness of using digital technologies in music education," *Journal of Educational Technology Development and Exchange (JETDE)*, vol. 17, no. 1, pp. 273–289, 2024, doi: 10.18785/jetde.1701.16.
- [7] M. Ouyang, "Employing mobile learning in music education," *Education and Information Technologies*, vol. 28, no. 5, pp. 5241–5257, 2023, doi: 10.1007/s10639-022-11353-5.
- [8] H. B. Yu, H. S. Hwang, D. H. Kim, H. J. Oh, and H. C. Kim, "Research on the Performance Optimization of HR-Net for Spinal Region Segmentation in Whole Spine X-ray Images," *Journal of Biomedical Engineering Research*, vol. 45, no. 4, pp. 139–147, 2024.
- [9] K. Ito, H. Nakashima, Y. Kagami, et al., "An appropriate method for predicting the femoral angle on whole-body X-ray images from full-spine X-ray images," *Journal of Orthopaedic Science*, vol. 29, no. 2, 2024, doi: 10.1016/j.jos.2023.02.007.
- [10] G. Yang, J. Lei, H. Tian, Z. Feng, and R. Liang, "Asymptotic feature pyramid network for labeling pixels and regions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 9, pp. 7820–7829, 2024, doi: 10.1109/TCSVT.2024.3376773.
- [11] D. Chen and S. Zhang, "Deep Learning-Based Involution Feature Extraction for Human Posture Recognition in Martial Arts," *Informatica*, vol. 49, no. 12, pp. 77–90, 2025, doi: 10.31449/inf.v49i12.7041.
- [12] Q. Ren, Z. Lu, H. Wu, J. Zhang, and Z. Dong, "HR-Net: A landmark based high realistic face reenactment network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 11, pp. 6347–6359, 2023, doi: 10.1109/TCSVT.2023.3268062.
- [13] H. C. Nguyen, T. H. Nguyen, J. Nowak, A. Byrski, A. Siwocha, and V. H. Le, "Combined YOLOv5 and HRNet for high accuracy 2D keypoint and human pose estimation," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 12, no. 4, pp. 281–298, 2022, doi: 10.2478/jaiscr-2022-0019.
- [14] Y. Wu, X. Chen, F. Dong, L. He, G. Cheng, Y. Zheng, et al., "Performance evaluation of a deep learning-based cascaded HRNet model for automatic measurement of X-ray imaging parameters of lumbar sagittal curvature," *European Spine Journal*, vol. 33, no. 11, pp. 4104–4118, 2024, doi: 10.1007/s00586-023-07937-5.
- [15] H. Zhang, Y. Dun, Y. Pei, S. Lai, C. Liu, K. Zhang, et al., "HF-HRNet: A simple hardware friendly high-resolution network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 8, pp. 7699–7711, 2024, doi: 10.1109/TCSVT.2024.3377365.
- [16] P. Lisena, A. Meroño-Peñuela, and R. Troncy, "MIDI2vec: Learning MIDI embeddings for reliable prediction of symbolic music metadata," *Semantic Web*, vol. 13, no. 3, pp. 357–377, 2022, doi: 10.3233/SW-210446.
- [17] S. Shen and K. Wu, "Solfeccio teaching method based on MIDI technology in the background of digital music teaching," *International Journal of Web-Based Learning and Teaching Technologies (IJWLTT)*, vol. 18, no. 1, pp. 1–18, 2023, doi: 10.4018/IJWLTT.331085.
- [18] E. D'Antonio, J. Taborri, I. Mileti, S. Rossi, and F. Patané, "Validation of a 3D markerless system for gait analysis based on Open Pose and two RGB webcams," *IEEE Sensors Journal*, vol. 21, no. 15, pp. 17064–17075, 2021, doi: 10.1109/JSEN.2021.3081188.
- [19] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 1, pp. 172–186, 2019, doi: 10.1109/TPAMI.2019.2929257.
- [20] A. del Barrio Mártel, "Wacom Intuos Pro 2025 Tablet: A Major Redesign," *FV: Photo-Video News*, vol. 38, no. 300, pp. 12–14, 2025, doi: 10.62077/qvoo62.mf858.
- [21] T. D. Nguyen, M. H. Shih, D. Srivastava, S. Tirthapura, and B. Xu, "Stratified random sampling from streaming and stored data," *Distributed and Parallel Databases*, vol. 39, no. 3, pp. 665–710, 2021, doi: 10.1007/s10619-020-07315-w.
- [22] R. S. Kenett, C. Gotwalt, L. Freeman, and X. Deng, "Self-supervised cross validation using data generation structure," *Applied Stochastic Models in Business and Industry*, vol. 38, no. 5, pp. 750–765, 2022, doi: 10.1002/asmb.2701.
- [23] Y. Wang, R. Wang, H. Shi, and D. Liu, "MS-HRNet: multi-scale high-resolution network for human pose estimation," *The Journal of Supercomputing*, vol. 80, no. 12, pp. 17269–17291, 2024, doi: 10.1007/s11227-024-06125-6.
- [24] M. Liu, K. Lin, W. Huo, L. Hu, and Z. He, "Feature enhancement modules applied to a feature pyramid network for object detection," *Pattern Analysis and Applications*, vol. 26, no. 2, pp. 617–629, 2023, doi: 10.1007/s10044-023-01152-0.
- [25] X. Song, J. Huang, J. Cao, and D. Song, "Feature spatial pyramid network for low-light image enhancement," *The visual computer*, vol. 39, no. 1, pp. 489–499, 2023, doi: 10.1007/s00371-021-02343-8.
- [26] G. Wang, X. Gan, Q. Cao, and Q. Zhai, "MFANet: multi-scale feature fusion network with attention mechanism," *The Visual Computer*, vol. 39, no. 7, pp. 2969–2980, 2023, doi: 10.1007/s00371-022-02503-4.
- [27] Z. Li, S. Lu, Y. Dong, and J. Guo, "Msffa: a multi-scale feature fusion and attention mechanism network for crowd counting," *The Visual Computer*, vol. 39, no. 3, pp. 1045–1056, 2023, doi: 10.1007/s00371-021-02383-0.
- [28] Y. Huang, P. Shi, H. He, H. He, and B. Zhao, "Senet: spatial information enhancement for semantic segmentation neural networks," *The Visual Computer*, vol. 40, no. 5, pp. 3427–3440, 2024, doi: 10.1007/s00371-023-03043-1.
- [29] C. Zeng, Y. Zhao, Z. Wang, K. Li, X. Wan, and M. Liu, "Squeeze-and-excitation self-attention mechanism enhanced digital audio source

- recognition based on transfer learning,” *Circuits, Systems, and Signal Processing*, vol. 44, no. 1, pp. 480-512, 2025, doi: 10.1007/s00034-024-02850-8.
- [30] S. Kılıçarslan, K. Adem, and M. Çelik, “An overview of the activation functions used in deep learning algorithms,” *Journal of New Results in Science*, vol. 10, no. 3, pp. 75-88, 2021, doi: 10.54187/jnrs.1011739.
- [31] C. Dewi, R. C. Chen, H. Yu, and X. Jiang, “Robust detection method for improving small traffic sign recognition based on spatial pyramid pooling,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 7, pp. 8135-8152, 2023, doi: 10.1007/s12652-021-03584-0.
- [32] J. Chen, F. Qin, F. Lu, L. Guo, C. Li, K. Yan, et al., “CSPP-IQA: a multi-scale spatial pyramid pooling-based approach for blind image quality assessment,” *Neural Computing and Applications*, vol. 37, no. 11, pp. 7499-7510, 2025, doi: 10.1007/s00521-022-07874-2.
- [33] M. Ju, J. Luo, Z. Wang, and H. Luo, “Adaptive feature fusion with attention mechanism for multi-scale target detection,” *Neural Computing and Applications*, vol. 33, no. 7, pp. 2769-2781, 2021, doi: 10.1007/s00521-020-05150-9.
- [34] J. Wang, H. Shao, Y. Peng, and B. Liu, “PSparseFormer: Enhancing fault feature extraction based on parallel sparse selfattention and multiscale broadcast feedforward block,” *IEEE Internet of Things Journal*, vol. 11, no. 13, pp. 22982-22991, 2024, doi: 10.1109/JIOT.2024.3377674.
- [35] A. Elen, S. Baş, and C. Közkurt, “An adaptive Gaussian kernel for support vector machine,” *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 10579-10588, 2022, doi: 10.1007/s13369-022-06654-3.