ITC 4/53
Information Technology and Control
Vol. 53 / No. 4/ 2024
pp. 1169-1187
DOI 10.5755/j01.itc.53.4.37391

Green Rural Modern Architectural Design Based on Pose Recognition Algorithm of Thermal Discomfort

Received 2024/05/22 | Accepted after revision 2024/08/02

HOW TO CITE: Gao, X. (2024). Green Rural Modern Architectural Design Based on Pose Recognition Algorithm of Thermal Discomfort. *Information Technology and Control*, 53(4), 1169-1187. https://doi.org/10.5755/j01.itc.53.4.37391

Green Rural Modern Architectural Design Based on Pose Recognition Algorithm of Thermal Discomfort

Xiaomei Gao
School of Design and Art, Shanghai Technical Institute of Electronics & Information, Shanghai 201411, China

Corresponding author: gracegao20212021@163.com

To reduce the energy waste of modern rural buildings caused by over-cold or over-heat supply, this paper presents a method to realize energy-saving design of modern green rural buildings by using thermal decomposition location recognition algorithm. Based on the key points of the human skeleton, the pose recognition framework is constructed, and the deep learning network is combined to detect the human thermal disturbance posture. Furthermore, an end-to-end thermal inauthentic pose recognition algorithm is proposed to establish a green intelligent building energy minimization model considering thermal comfort range. The results show that the recognition rate of 1D convolution +LSTM model is 100%, and the optimal accuracy of 16 frames of image sequence with decoder module is 92.052%. Compared to traditional algorithms, the method can save up to 10% of the total energy cost and reduce the total temperature deviation. This study is of great significance for intelligent control of indoor thermal environment and improvement of energy utilization efficiency.

KEYWORDS: Thermal discomfort; Pose; Recognition; Rural areas; Architecture; Energy saving.

1. Introduction

People's demand for comfort in life is increasing, which is accompanied by increasing energy consumption. Currently, about 20% of the world's energy consumption comes from buildings, with half of it coming from Heating, Ventilation and Air Conditioning (HVAC) systems. The main method of climate change framework is to provide a relatively constant temperature and humidity environment for buildings in accordance with relevant international standards [8, 21]. Housing construction in rural areas is usual-

ly carried out around agricultural development, but in the construction of new rural areas, some village planning does not fully consider meteorological factors such as sunshine and ventilation, and the layout is random. For example, the east-west orientation of residential areas in some villages in order to provide rental shops along the road leads to excessive indoor temperatures in the summer and the need to use air conditioning for a long time, resulting in wasted electricity. With the improvement of rural economy and residents' living standard, centralized residential areas are being built in various rural areas. At present, the main goal of energy conservation in rural buildings in China is to reduce energy consumption while ensuring the thermal comfort of occupants [1]. Existing studies mainly use contact and semi-contact methods to detect Thermal comfort (TC), but these methods are dedicated to equipment and invasive to the tested [9]. Questionnaire survey can obtain users' subjective feelings, but the efficiency is low. Environmental monitoring methods can obtain thermal comfort sustainably, but with low accuracy. The physiological parameter detection method is more accurate, but it requires contact with equipment [18]. To overcome these shortcomings, this paper proposes a camera-based non-contact method to design thermal uncomfortable pose recognition algorithm through machine vision and deep learning technology. This algorithm aims to realize the green Energy saving (ES) design of modern rural buildings, accurately obtain the environmental thermal comfort, and avoid energy waste. Through the construction of green intelligent building energy cost minimization model, it is expected to achieve intelligent regulation of indoor thermal environment, improve energy efficiency, and support the promotion of green environmental protection concept.

The contributions of the research are as follows: First, a thermal uncomfortable pose recognition algorithm based on bone key points is proposed, and an end-to-end detection system is designed. Deep learning technology is used to improve the recognition accuracy and efficiency. Secondly, based on this algorithm, the energy cost minimization model of green intelligent buildings is established, and the combination of thermal comfort and energy saving requirements is realized to achieve dynamic energy management of intelligent buildings. In addition, the non-contact detection method overcomes the limitations of traditional methods and improves the convenience and user experience. Finally, the proposed algorithm is superior to the traditional method in accuracy and energy saving effect, and can save energy significantly under the condition that the total temperature deviation is small.

The research mainly includes five parts, and in the first part of the article, the background and significance of research on human TC detection are mainly introduced. The content of Part 2 is a comprehensive overview of educational games, mainly focusing on a detailed analysis of the achievements of experts and scholars at home and abroad in the field of TC detection. The third part is the research methodology, mainly divided into two sections. In section 1, the study proposes a TDPRA based on bone key (BK). In the second section, to further improve PR accuracy, an end-to-end TDPRA was proposed. The fourth part is about verifying the effectiveness of the research model. The fifth part is a summary of the most research methods and an analysis of experimental results. And the shortcomings of research methods and future research directions are proposed.

2. Related Works

In the total energy consumption, building energy consumption accounts for a large proportion and has become a major issue of concern to society. A comfortable environment has a positive regulation on human physiological parameters, and ES and TC are closely related to each other, which has also sparked in-depth exploration by many researchers both domestically and internationally. Meng et al. [16] proposed AdaViT, a visual deformer based on a self-attention mechanism, aimed at reducing the computational cost of visual tasks. AdaViT improves inference efficiency by learning to adaptively adjust the strategy of using patches, self-focusing heads, and transformer blocks for each image. The framework attaches a lightweight decision network to the transformer backbone to generate real-time decisions and perform end-to-end optimization. In the ImageNet experiment, AdaViT achieved a more than two-fold increase in efficiency with only a 0.8% reduction in accuracy, balancing efficiency and accuracy. Ramsey et al. [18] studied the

application of human energy consumption analysis methods in indoor thermal environment evaluation and established a new dual node energy consumption analysis model. The rate of human exergy loss has an extreme value at low or high working temperatures, and can be used separately to evaluate human TC resistance. Chen et al. [7] studied how to improve image classification performance by learning multi-scale feature representation in transformer models. For this purpose, a double-branch transformer is proposed, which combines image patches of different sizes to generate stronger image features. This approach uses separate branches to handle small patches and large patch tokens, complementing each other through multiple attention fusions. In addition, a token fusion module based on cross-attention is developed, requiring only linear time computation and storage complexity. Experiments show that the proposed method performs better than DeiT on ImageNet1K dataset, and FLOPs and model parameters are slightly increased [7]. Scholars such as Qabbal et al. [17] use specially developed intelligent sensors to detect air pollutants and TC levels inside buildings. Most residents feel uncomfortable with indoor temperature [25]. Kong et al. [13] compared the TC performance and energy efficiency of eight widely used space heating and ventilation methods. The TC air inlet is preferably located at a high horizontal position in the wall, while the air outlet is located at the same or higher height position [13]. Jaffal et al. [10] designed a non-airconditioned building TC metamodel based on physical information machine learning to improve the interpretability and accuracy of the building TC machine learning model. This model combines the advantages of physics and machine learning, and can support architectural design with flexible and interpretable metamodels. Traffic sign recognition is the key task in automatic driving. Zheng et al. [26] propose a camera-based computer vision technique that uses multiple convolutional neural network structures and is validated on multiple datasets. Recently, new transformer-based models have outperformed convolutional neural networks in a variety of vision tasks. However, the study in [11] found that transformers did not perform as well as convolutional neural networks in traffic sign classification tasks.

Researchers such as Bia and Koltuk [5] provided a detailed introduction to the TC measurement method of the modern intelligent building "Energis", and analyzed the indoor air parameters and subjective reactions of volunteers. People feel better in environments where TC is considered [5]. Tummala et al.'s [20] Vision transformer-based (ViT) deep neural networks have gained a lot of attention in the field of computer vision due to their success in natural language processing. The study explored the efficacy of the ViT model in the diagnosis of brain tumors in T1-weighted (T1w) magnetic resonance imaging. The pre-trained and fine-tuned ViT model on ImageNet was used to classify MRI sections of brain tumors. The best model L/32 has a test accuracy of 98.2% at 384 × 384 resolution, and the accuracy of the four ViT models integrated is 98.7%, which is higher than the performance of a single model at 224 × 224 resolution [15]. Kwong et al. [14] analyzed the distribution of TC parameters such as air temperature and velocity. And electronic sensors are used to collect the information needed for the prediction of air temperature and velocity distribution mode in laboratory and workshop based on Fluid mechanics model. Experiments have shown that this model not only helps to design efficient air conditioning and mechanical ventilation (ACMV) systems, but also helps to improve indoor TC performance [14]. Scholars such as Austin [4] proposed the design process of a prototype Air Port Controller Module (PCM) device and provided a detailed introduction to the corresponding experimental testing. The unit limits the indoor air temperature rise during operation, keeping the temperature within the TC range, thereby helping to reduce thermal discomfort [4].

In summary, most research is focused on one field or technology, such as vision transformation models and smart sensors, and there is a lack of integrated methods for building energy consumption and TC detection. Many studies focus only on specific scenarios, such as indoor environments or specific types of buildings, ignoring extensive validation in different types of buildings and environments. Although some studies attempt to improve model explainability, model transparency and explainability remain challenges in complex environments, especially in applications where machine learning and physical models are combined. Data collection and processing still have limitations in terms of long-term and large-scale monitoring, and real-time and accuracy need

to be improved. Therefore, the study conducted in-depth analysis of non-contact physiological detection methods and proposed an ES design based on TDPRA for green rural modern architecture (GRMA).

3. GRMA Design Based on TDPRA

Approximately 20% of the world's energy consumption comes from buildings, with 50% of building energy consumption coming from HVAC systems. Visual perception technology can perform non-contact detection of human TD, effectively solving the energy recovery problem caused by inaccurate cold and hot supply. Therefore, the study first designed a framework for TDPRA based on human BK. On this basis, to improve its detection accuracy, an end-to-end human TDPRA was proposed. Finally, based on TDPRA, a rural household HVAC system model was constructed. By utilizing its passive regulation to ensure indoor users' TC, a green intelligent building design has been achieved.

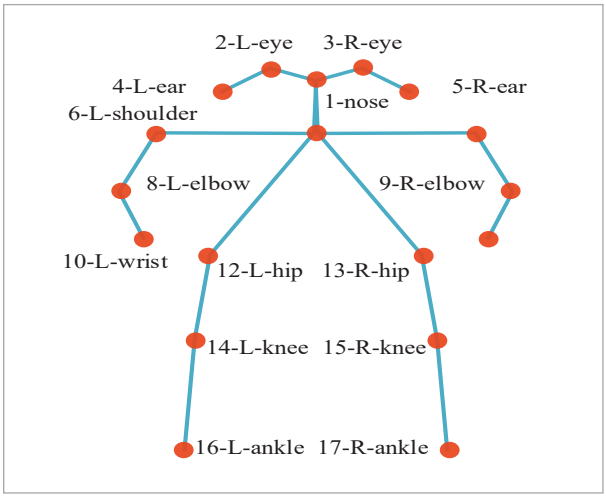
3.1. Design of TDPRA Framework Based on BK

A good human TC environment plays a crucial role in the ES of intelligent buildings and personal health. In the early 20th century, Hill proposed that the temperature and airflow inside buildings should be designed based on the human body's TC needs [24]. The temperature, air flow rate, and humidity in buildings have varying degrees of influence on indoor TC degree. Most human bodies express TD through body language. The relationship between body language and Thermal Comfort (TC) is determined by observing the posture and movements of the human body in different temperature environments. This helps smart building systems automatically adjust the indoor environment based on human posture, improving comfort and saving energy. For example, heat discomfort includes actions such as wiping sweat, fanning, shaking clothes, scratching your head, and rolling your sleeves. Comfort includes normal posture such as standing and walking. Cold discomfort is manifested as arms, crossed legs, stamping feet, rubbing hands and other actions. Therefore, a corresponding relationship between body language and TC status can be established. Due to the bottom-up estimation of human posture, the first step is to detect human skeletal nodes. Therefore, this method inevitably includes candidate key points and BK detection [12]. The extraction module of BK belongs to the category of 2D attitude estimation, which adopts the form of the COOC2017 dataset. Figure 1 shows a dataset-based graph of 17 BKs in the human body.

In Figure 1, a BK annotation contains all the data of the object. BK is an array of length 2×K, and K is the total BKs defined for that category. COOC2017 dataset used in the study provided 17 BKs, with a BK array length of 34. 17 BKs are connected to the trunk by 19 connecting lines, corresponding to human joints such as the nose, eyes, ears, shoulders, elbows, wrists, buttocks, knees, and ankles. Research will use the Discontinued Key Regression (DEKR) algorithm as the extraction module for BK [19].

Dense key point regression will represent one pixel

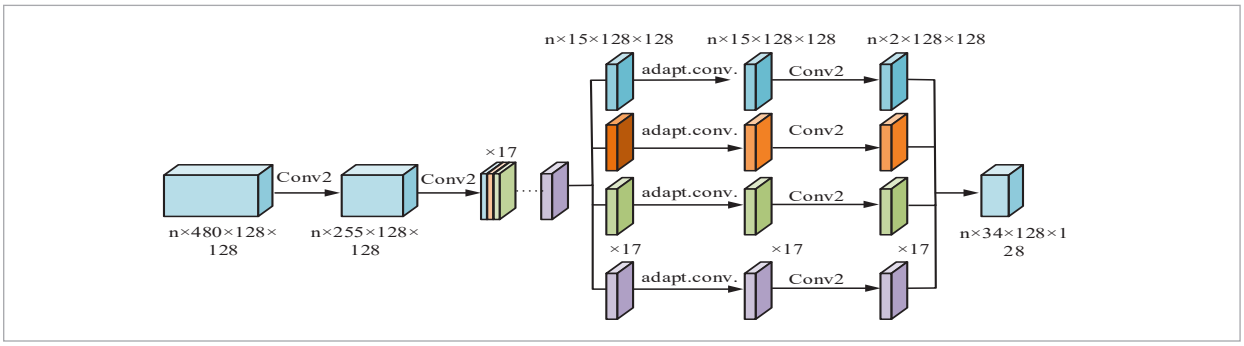
Figure 1
Human joints corresponding to key points of bone



with you through an offset vector with a dimension of 2k, thereby achieving pose estimation. Figure 2 shows the DEKR framework structure.

In Figure 2, DEKR structure's output is HRNet's output, and there are 17 key points in human pose estimation for COCO data. Therefore, DEKR involves regressing the coordinates of 17 key points separately. DEKR consists of two committed steps. Firstly, adaptive convolution was used to activate pixels near key points, and depth features were learned based on the activated pixels. Adaptive convolution's kernel is ob-

Figure 2
DEKR frame structure



tained from the predicted affine transformation matrix and translation matrix generated by each pixel, as shown in Formula (1) [23].

$$\begin{cases} G_s^q = A^q G_t + [tt\ldots t] \\ G_t = \begin{bmatrix} -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \end{cases} \quad (1)$$

In Formula (1), A^q stands for the Affine transformation matrix. G_t stands for 3 × 3 ordinary two-dimensional convolution. $[tt\ldots t]$ stands for the convolution of the input dimension. And mapping the number of channels to 2 yields $[2, H, W]$. Among them, [2] stands for the coordinates on the x and y axes, and [9] stands for 3 × 3 offset positions. G_s^q stands for the final offset. q stands for x and y axes' center in Equation (2).

$$y(q) = \sum_{i=1}^{9} W_i x\left(g_{si}^q + q\right). \quad (2)$$

In Equation (2), g_{si}^q stands for the offset of two-dimensional coordinates, and W_i stands for the convolutional kernel weight matrix. Multi branch regression with the same branch structure uses adaptive convolution to obtain feature information of pixels around the center for each branch, and the coordinate information of key points is obtained by regression in Equation (3).

$$\begin{cases} O_1 = F_1(X_1) \\ O_2 = F_2(X_2) \\ \vdots \\ O_K = F_K(X_K) \end{cases} \quad (3)$$

In Formula (3), X is the output eigenvector of the backbone network, and Y stands for a single same score structure obtained by independent training. K stands for the BKs number, which is 17. PR module extracts BK and obtains the coordinates of key points with a dimension of n×17×2. After normalizing the coordinates, the data is then processed through the pose recognition algorithm (PRA) to achieve attitude recognition. Where n is the frames in the action sequence taken by PRA input. BK based PR belongs to multi classification tasks, so its main focus is on the prediction accuracy of classification. The cross entropy loss function is used as this algorithm's loss function in Formula (4) [28].

$$CEL(W) = \frac{1}{N} \sum_{i=1}^{N} H\left(y^{(i)}, \hat{y}^{i}\right). \quad (4)$$

In Equation (4), W stands for the network model parameter matrix. N stands for the number of training samples. $y^{(i)}, y$ represent the true and predicted values of label, respectively. $H\left(y^{(i)}, \hat{y}^{i}\right)$ stands for the information entropy of $y^{(i)}, y$, which is specifically expressed in Equation (5).

$$H\left(y^{(i)}, \hat{y}^{i}\right) = -\sum_{j=1}^{q} y_j^{(i)} \log \hat{y}_j^{(i)}. \quad (5)$$

In Equation (5), $y_j^{(i)}$ stands for the element in $y^{(i)}$ which is 0 or 1, and the correct category is 1. Therefore, cross entropy can be transformed into Equation (6).

$$H\left(y^{(i)}, \hat{y}^{i}\right) = -\log \hat{y}^{y^{(i)}}_{y^{(i)}}. \quad (6)$$

In Equation (6), $y^{(i)}$ is the true category label of class i. $\hat{y}^{(i)}_{y^{(i)}}$ stands for softmax function's output in Equation (7).

$$\hat{y}_k = soft\max\left(o_k\right) = \frac{\exp\left(o_k\right)}{\sum_{i=1}^{q}\exp\left(o_i\right)}. \quad (7)$$

In Equation (7), $\hat{y}_k$ stands for softmax function's output $\hat{y}^{(i)}_{y^{(i)}}$. o_k is neural network's output. The softmax function converts network's output into a probability value between 0 and 1. Accuracy is used in both training and testing stages. It stands for the accuracy of model's prediction result in Equation (8).

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} == \hat{y}^{(i)}\right). \quad (8)$$

In Equation (8), N stands for the number of samples, and $y^{(i)} == \hat{y}^{(i)}$ stands for the equality of predicted and true values. The predicted value was counted as equal to true value. Then the proportion of total sample size is accuracy. After obtaining BK coordinate information, the features input to network are relatively less than image information. Therefore, three algorithms, namely full connection network, 1D convolution network and 1D convolution+Long short-term memory (LSTM) network, were constructed for attitude recognition [22]. The matrix dimension of fully connected network input data is n×5×34, with a feature vector length of 5×34. Figure 3 shows a fully connected network [6].

In Figure 3, the input data dimension is n×5×34. After Reshape, the dimension of its feature vector is n×1×170. The feature vector dimension obtained through the fully connected Densel structure is n×1×256, followed by Dense2, Dense3, Dense4, and Dense5 structures to obtain n×1×64 dimensions. Finally, through the final layer of fully connected structure Dense6, the feature vector dimension is n×1×17. The fully connected structures' dropouts above are all 0.5, and then the feature vector dimension obtained through Flatten is n×17. At this point, the output of 17 stands for 17 posture categories. The input data dimension for 1D convolutional networks is n×5×34, the length of the feature vector is 5×34. Figure 4 shows the network structure [15].

Figure 3
Structure of a fully connected network

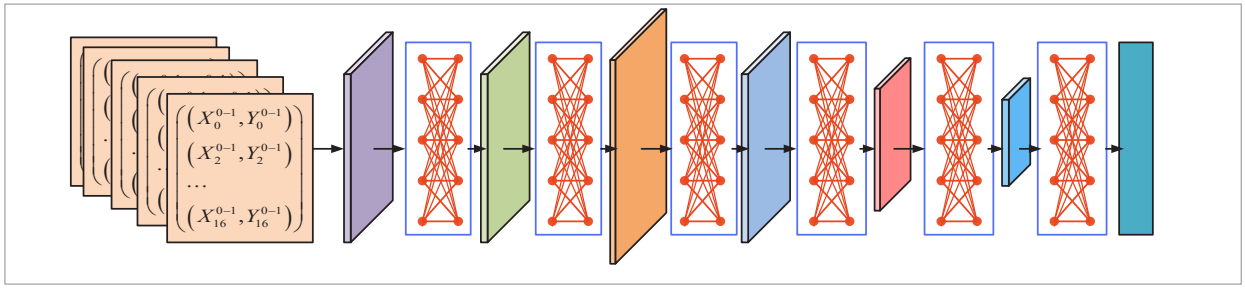
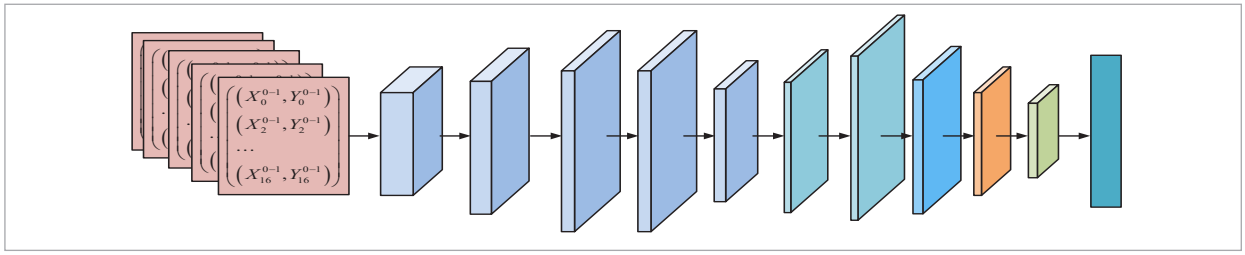


Figure 4
Structure of 1D convolutional network



In Figure 4, the dimension of the input data is first, and the dimension of the feature vector obtained through Reshape is n×10×17. Then, after passing through the 1D convolutional structure CONV in sequence, the dimension of the feature vector is n×6×32. Then, after Flatten, the dimension of the feature vector is n×192. The dimension of the feature vector obtained through a 5-layer fully connected structure. The Dropouts of the fully connected structure are all 0.5, and at this time, the output of 17 stands for 17 posture categories. The structure of the 1D convolution+LSTM algorithm has two branch structures, namely the LSTM branch and the 1D convolution branch. The data dimension n×5×34 was first input. Firstly, the 1D convolution input is processed through Reshape to obtain a feature vector with a dimension of n×10×17. Then, through a four-layer 1D convolution structure, the feature vector dimension is obtained as n×6×32. After passing through Flatten, the feature vector dimension is n × 192. Through structures LSTM_1 to LSTM_5, the feature vectors dimensions were obtained as n×5×32, n×5×64, n×5×128, n×5×64 and n×96. Then two branches' feature vectors are fused to obtain a feature vector dimension of n×288. n×17 is obtained through a 5-layer fully connected structure. The fully connected structure's dropouts are all 0.5, and at this time, the output of 17 stands for 17 posture categories.

3.2. Design of Rural Green Intelligent Buildings Based on TDPRA Detection

Due to the interference caused by similar attitudes in BK coordinate data, the difference in normalized similar attitude coordinate data is smaller, resulting in insufficient similarity PR. To improve detection accuracy, an end-to-end form has been proposed for TD attitude recognition. Residual Net (ResNet) effectively solves the problem of performance degradation when there are too many layers in deep learning networks. The introduction of residual structures provides feasibility for extracting more complex feature information and improving network performance [2]. Based on end-to-end TDPRA, motion sequence images are used as input to recognize and detect human thermal discomfort postures by extracting temporal relationships between motion sequences and image features. A dataset was constructed for the research algorithm, with input from the network consisting of video sequences of 8 and 16 frames. An average of 8 or 16 frames of image data were selected from an action as the input algorithm representing the complete action. There are a total of 17 categories based on this action, and Equation (9) is the calculation of accuracy.

$$Accuracy@1 = \frac{1}{N}\sum_{l=1}^{N}\left(y^{l} == \hat{y}^{(l)}\right). \tag{9}$$

In Equation (9), N stands for the number of samples, and $y^{l} == \hat{y}^{(l)}$ stands for that the predicted value is equal to true value. The traditional Resnet50 input matrix's dimension is (3, 224, 224), which correspond to the channel number, height, and width, respectively. The research is based on a classification method for video sequences, so this algorithm's input matrix dimension is (3 * 8/, 224, 224). That is, 8/16 frames of images containing the entire action information are superimposed as channels for the entire network. Figure 5 shows the ResNet50 network.

In Figure 5, ResNet50 goes through five stages. The first stage involves preprocessing operations, while the remaining four stages involve four layers of convolution. The number of bottleneck modules corresponding to each convolution layer is 3, 4, 6, and 3, respectively. The bottleneck module has performance limitations in modeling the time dimension. By adding a Motion Excitation (ME) module to the bottleneck module for stimulating motion patterns, as well as a Multiple Temporal Aggregation (MTA) module for establishing long-range time relationships, time dimension modeling can be achieved [3]. 3×3 convolution in the bottleneck module was replaced with a combination of ME and MTA. The temporal excitation and aggregation (TEA) module is composed of 1×1 convolution before and after combining [28] Figure 6 shows the network structure based on TEA module.

In Figure 6, an average of 8/16 frames are selected from video frame sequence as input to network. Through five stages, the first stage is to preprocess the input data, and the latter step is also a convolutional process composed of several TEA modules. Finally, the feature information output from four layer convolution is passed through the fully connected layer, and 17 action types output by network are converted into output probabilities through softmax function. Finally, the one with the highest probability was out-

Figure 5
ResNet 50 network structure

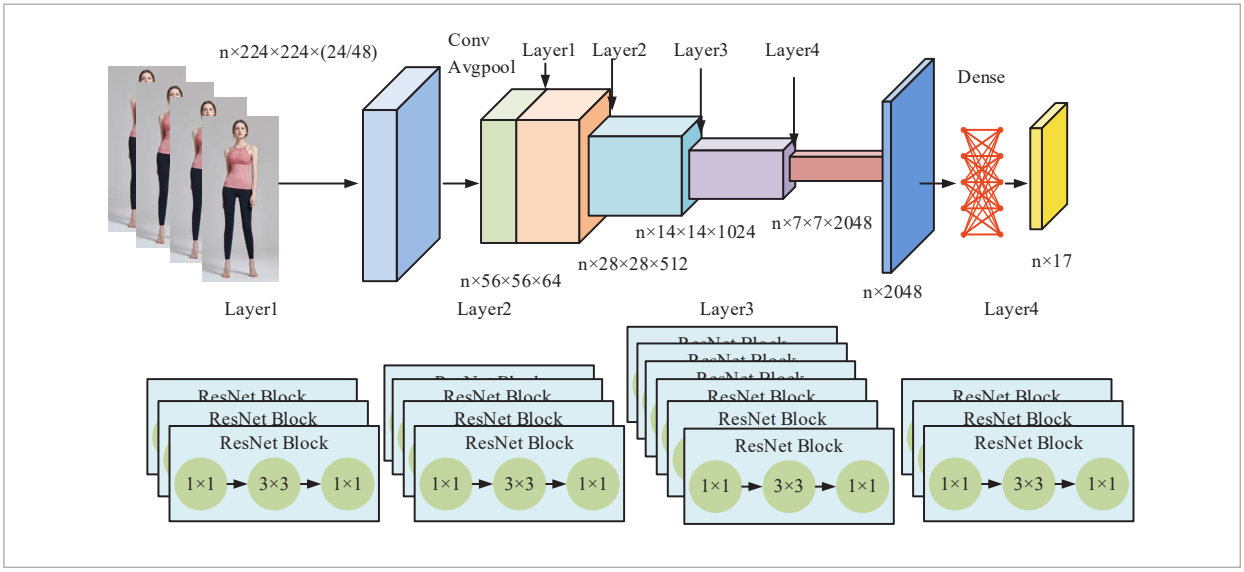
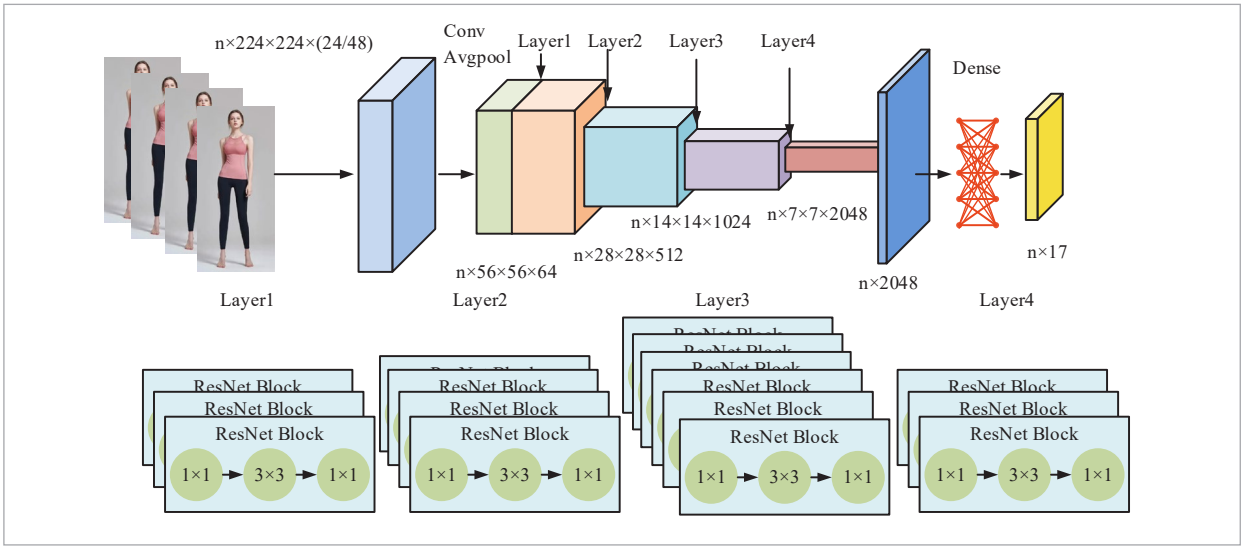


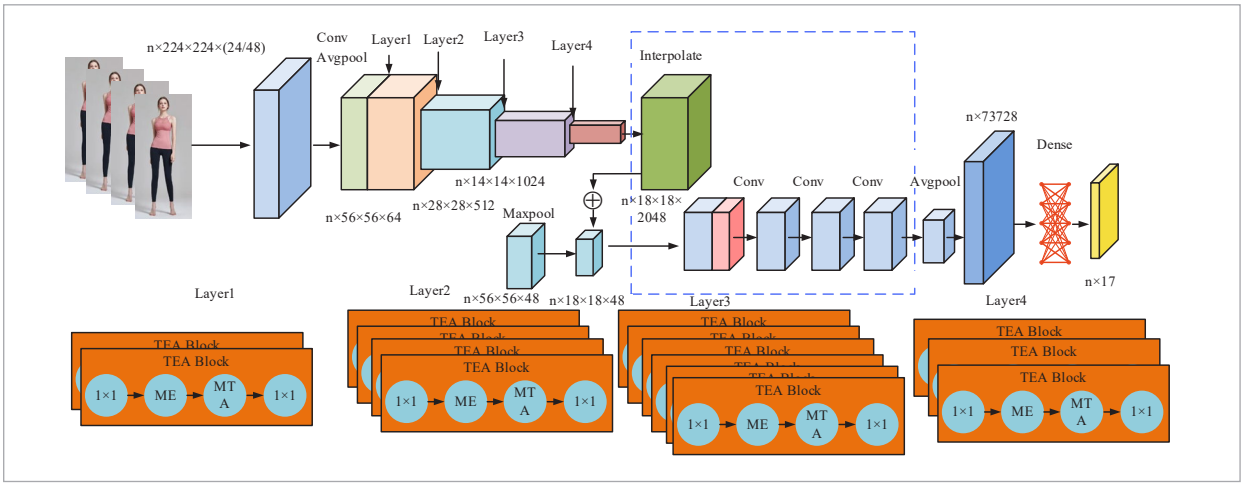
Figure 6
Network structure based on TEA module



put as the result. In the last four stages, the output characteristic dimension and reaction network 50 remain unchanged. On this basis, this article focuses on the application of MT and MTA models in the time domain. After implementing modeling in the temporal dimension, the study combines low-level features with high-level features and proposes a PRA based on TEA and Decoder. Figure 7 shows the network structure based on TEA and Decoder.

In Figure 7, the network structure is based on PR network structure on the foundation of TEA, encoding the feature information output from the convolution of the first, second, and fourth layers. The decoded feature information is fully connected to obtain the

Figure 7
Network structure based on TEA and Decoder



action category. When the convolutional layer increases, hidden features are easily extracted. However, increasing layer number will result in more expression of high-level features, while low-level features will be missed due to convolution, which will have an impact on the results. The decoding module constructed through research can effectively integrate high-level and low-level feature information, and use high-level and low-level feature information as the final fully connected input. This achieves a balance between high and low level feature information, thereby improving the accuracy of pose recognition. Decoder first transforms the underlying feature vectors into 1×1 through one convolutional kernel, reducing the number of channels to 48. The second step is to linearly interpolate the high-level feature information, so that the dimension of the low-level feature information is equal to that of the high-level feature information. The third step is to fuse the high-level and low-level feature vectors after Linear interpolation. Finally, the fused feature vectors are output through the final convolution layer and used as inputs to the fully connected layer. In the design of green smart buildings, indoor HVAC systems can be dynamically adjusted to ensure user TC. Equation (10) stands for the comfortable temperature range to represent the user's TC.

$$T^{\min} \le T_t \le T^{\max}, \forall t. \tag{10}$$

In Equation (10), $T^{\min}, T^{\max}$ represent the lower and upper limits of indoor comfortable temperature, which can be set according to the needs of user TC. Considering the variable frequency function of the HVAC system, its input power e_t can be continuously adjusted. If the rated power of HVAC system is e_t, there is Equation (11).

$$0 \le e_t \le e^{\max}, \forall t. \tag{11}$$

In the energy management system of intelligent buildings, it needs ensure that entire energy management system's total power supply is equal to the total required power supply. Thus, it can ensure the power supply balance in energy management system. The calculation expression is shown in Equation (12).

$$g_t + p_t - d_t = b_t + e_t + c_t, \forall t. \tag{12}$$

In Equation (12), g_t, p_t, b_t represent the power output from the large power grid, renewable energy generation, and rigid load demand power, respectively. c_t, d_t represent the charging and discharging power of energy storage system. If $g_t \ge 0$, rural smart building households will purchase electricity from large power grids. Otherwise, it will sell the excess electricity to the large power grid. Assuming that v_t, u_t are the electricity prices purchased and sold by rural intelligent building households, Equation (13) is the energy cost of time slot t participating in the purchase and sale.

$$C_{1,t}=\left(\frac{v_t-u_t}{2}\left|g_t\right|+\frac{v_t+u_t}{2}g_t\right),\forall t. \tag{13}$$

In Equation (13), it stands for the behavior of rural smart building households buying or selling electricity using only one variable g_t. Frequent charging and discharging can reduce the service life of energy storage systems, so the depreciation cost of time slot t energy storage systems is defined as Equation (14).

$$C_{2,t}=\left(\psi\left|c_t\right|+\left|d_t\right|\right),\forall t. \tag{14}$$

In Equation (14), ψ stands for energy storage system's depreciation coefficient. Based on these algorithms above, a green intelligent building energy cost minimization model considering the TC range was established in Equation (15).

$$\begin{cases} B_{t+1}\min\sum_{t=1}^{T}\mathrm{E}\left\{C_{1,t}+C_{2,t}\right\} \\ s.t.B_{t+1}-g_t-p_t+d_t \end{cases}. \tag{15}$$

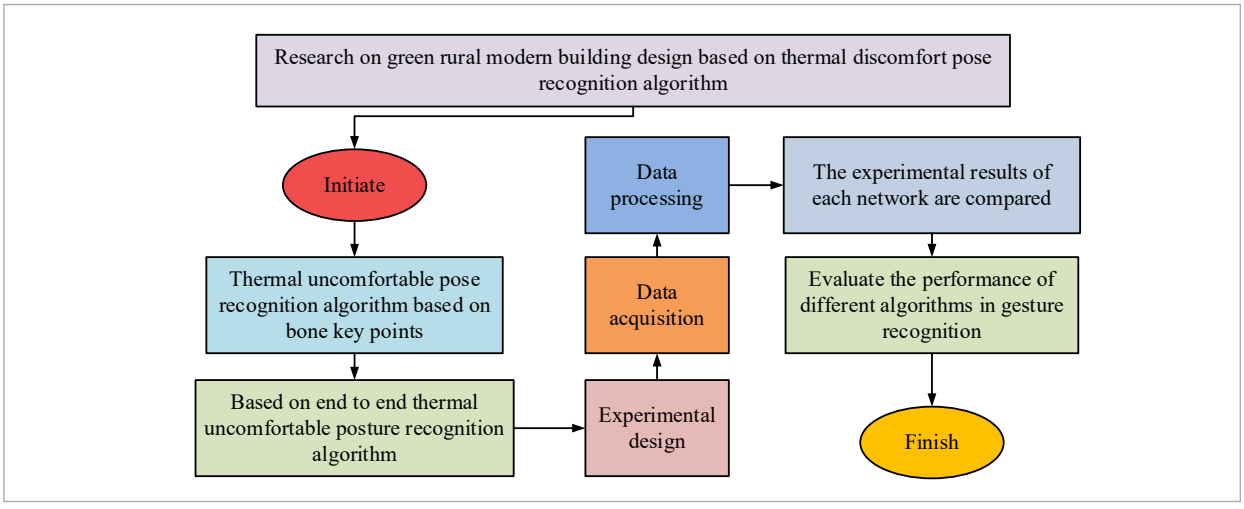
In Equation (15), B_{t+1} dynamically stands for the energy storage system's energy storage level in time slot t. This expected operation affects the randomness of system parameters and possible control behaviors for each time slot uncertainty. Figure 8 shows the technical framework of green rural modern building design research based on thermal discomfort pose recognition algorithm.

In Figure 8, the research methods include two thermal discomfort pose recognition algorithms based on bone key points and end-to-end. In the former, 17 skeleton key points from COOC2017 dataset were extracted by DEKR algorithm, and 1D convolutional network and 1D convolutional +LSTM were used for pose recognition, and accuracy and loss values were evaluated. The latter is identified using a combination of deep residual network (ResNet), motion excitation (ME), and multi-time aggregation (MTA) modules, optimized with TEA modules and Decoder. The experiment involved computers and high-definition cameras, 18 volunteers performing 17 thermal uncomfortable poses, generating 8,000 videos and 320,000 images, the data for training and testing, recording loss values and accuracy, and comparing the performance of different models. The optimized algorithm is used in HVAC system of green intelligent building to ensure indoor thermal comfort and achieve energy saving through passive adjustment.

4. Performance Analysis of GRMA Design Model Based on TDPRA

The experimental device includes one computer and one high-definition USB camera. The height of the camera from the ground is 1.5m. Before conducting action collection, the experimental object must first

Figure 8
Research technology framework



reach a designated area 1.7m away from the camera to ensure that all actions of the experimental object can be captured by the camera. Due to insufficient data in attitude-based TC detection direction, a dataset based on TD attitude was constructed. This dataset invited a total of 18 adults aged 18-28 as experimental subjects, each of whom collected 17 different TD postures. There are two types of data formats: video and image, with around 8000 video data and around 320000 images. Human body's different postures represent the cold and hot states they are in. In addition to accuracy, precision, recall rate and cross entropy loss are also used to evaluate the performance of the thermal uncomfortable posture recognition algorithm. Accuracy refers to the proportion of the number of samples correctly predicted for a certain class to the number of all samples predicted for that class; The recall rate is the proportion of the number of samples correctly predicted for a category to the actual number of samples for that category. Cross entropy loss is used to measure the difference between the model's predicted probability distribution and the true distribution. Table 1 shows the meanings and explanations of 17 movements and postures.

In Table 1, walking indicates that the human body is in a normal thermal environment, while the other 16 indicate TD posture. The experiment conducted 3000 rounds of iterative training, with each round corresponding to one test. In the 1D convolution layer, the convolution kernel size is 3, the number of convolution nuclei is 32, the step size is 1, and the filling method is the same to capture the local features of the se-

Table 1
Postural description and sample quantity of 17 actions

Serial number	Attitude name	Implication	Attitude declaration	Sample size
1	Raise hands to wipe sweat	Hot	Put your hand on your forehead to wipe the sweat	512
2	Raise your hand and fan	Hot	Put your hands next to your head and fan	489
3	Shake clothes	Hot	Put your hand on your chest and shake your clothes	447
4	Raise one's hand and scratch one's head	Hot	Put your hands on your head and scratch your head	378
5	Roll up one's sleeves	Hot	Pull the cuffs of the clothes in with your hands. Pull up	562
6	Open position	Hot	Open hands	456
7	Head wave	Hot	Hands on head. Air fan	853
8	Lapel	Hot	Hand pulling the side collar of the neck	469
9	Walk	Hot	Walk normally	456
10	Standing shrug	Moderation	Shoulders up, head down	472
11	Fold one's arms	Cold	Cross your hands over your chest	652
12	Cross leg	Cold	Cross left leg and right leg	457
13	Hand on neck	Cold	Put your hands on your neck	498
14	Raise one's hand to vent	Cold	Put your hands together in front of your mouth and nose	561
15	Stomp on the ground	Cold	Step on the floor with your left and right feet alternately	473
16	Lift and rub hands	Cold	Rub your hands alternately in front of your abdomen	421
17	Contraction position	Cold	Bend your arms and pull your neck down	470

quence. LSTM layer parameters include the number of hidden units (such as 32, 64, 128) and the number of layers (1 to 3 layers), dropout of 0.5 is used in training to prevent overfitting, the number of samples per training is 16, 32, or 64, the optimizer selects Adam, and the learning rate is 0.01. During the training of a fully connected network, changes in loss values and accuracy on both the training and testing sets are recorded in Figure 9.

According to Figure 9(a), training set's loss value decreases with cycles increasing and significantly decreases between 0-1800 cycles. As training increases, the loss value gradually decreases. In addition, the accuracy continuously improves with the iteration. At 0-1850 rounds, the accuracy will rapidly improve, then slowly increase, and ultimately reach 85.18%. According to Figure 9(b), test group's accuracy decreased significantly from 0 to 900 rounds, and then stabilized. The prediction accuracy has significantly improved between 0 and 950, and then stabilized, reaching the optimal value of 67.99%. When training a 1D convolutional network, Figure 10 shows the changes in loss values and accuracy on different sets.

According to Figure 10(a), training set's overall loss values show a trend of being taken off the shelves, while the accuracy rate shows an overall upward trend. Both showed significant changes in the 0-1200 rounds of data, followed by slow changes, with an ac-

Figure 9

Variation trend of loss value and accuracy on training set and test set of fully connected network

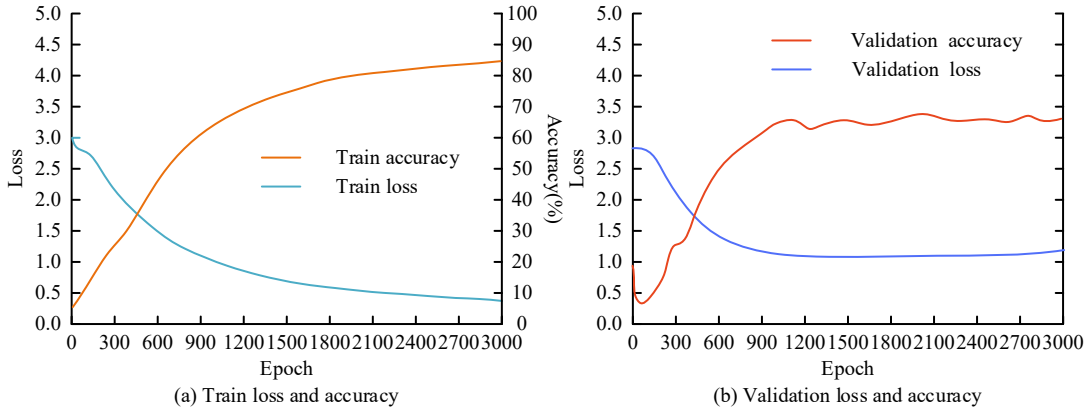


Figure 10

Variation trend of loss value and accuracy on training set and test set of 1D convolutional networks

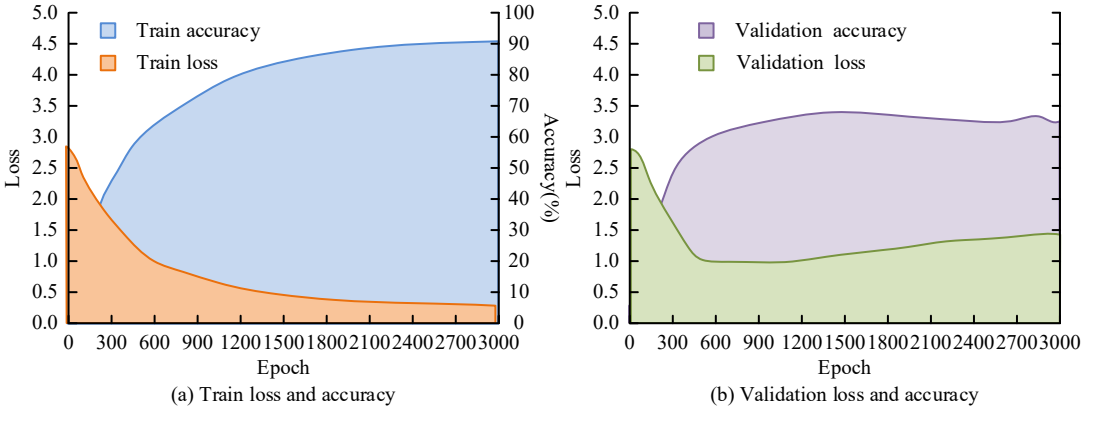
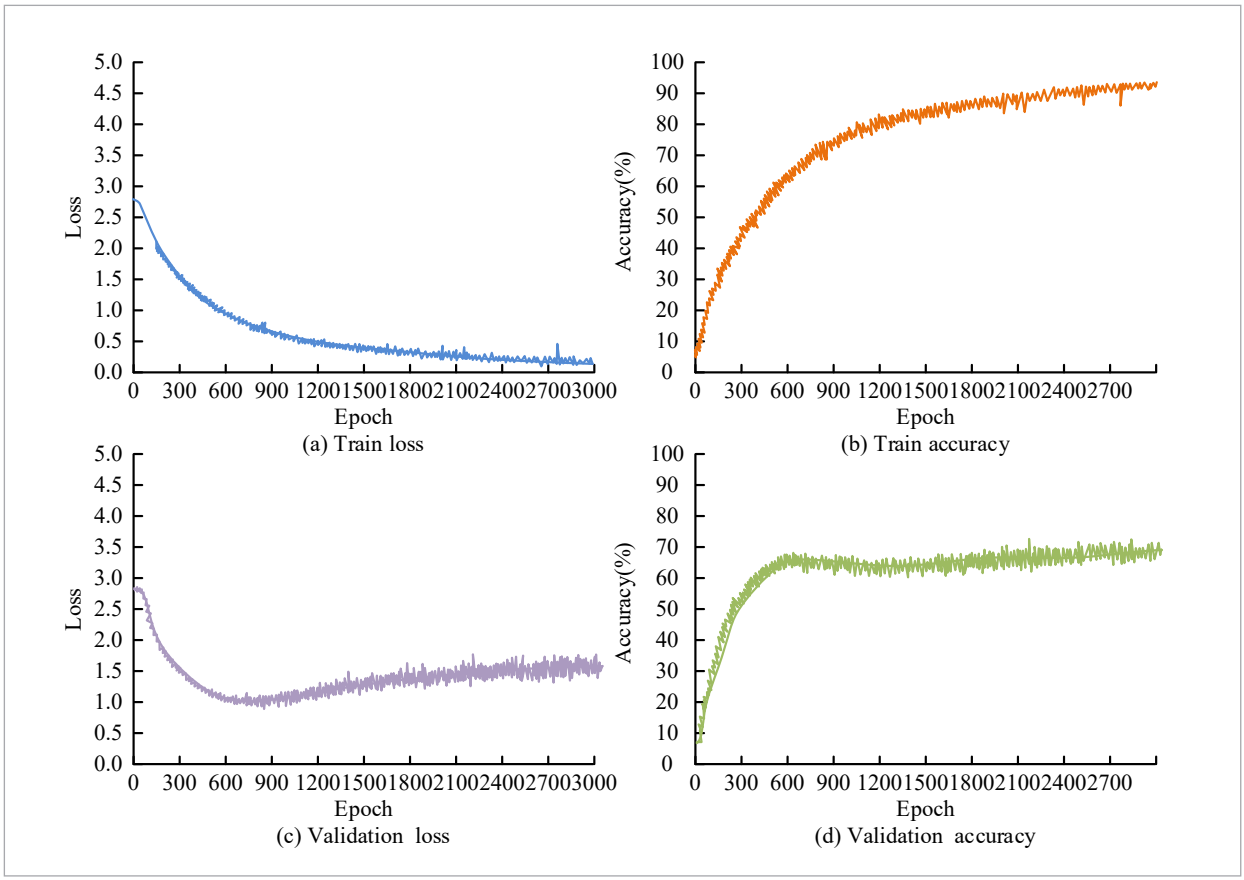


Figure 11
Variation trend of loss value and accuracy on training set and test set of 1D convolutional +LSTM convolutional network

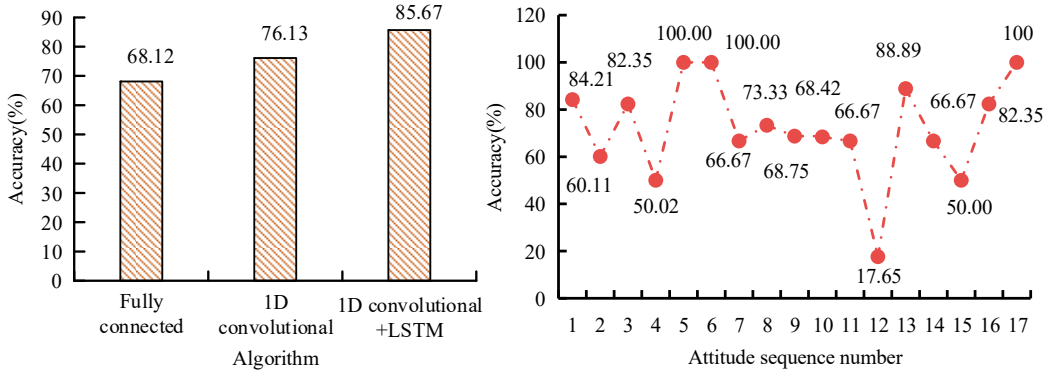


curacy rate of over 91.12%. According to Figure 10(b), the loss value on the test set reaches the lowest level between 600 and 1200 rounds, with the highest accuracy reaching 71.23%. Afterwards, the fitting loss value reached its lowest point and began to rise, while the accuracy began to decline. When training 1D convolution+LSTM network, Figure 11 shows the changes in loss values and accuracy on different datasets.

According to Figure 11(a), the overall loss value on the training set shows a decreasing trend, with a relatively rapid convergence in the 0-600 rounds and a slower trend afterwards. According to Figure 11(b), the rate of increase in accuracy is the fastest between 0-600 rounds, and the final frame is above 92.09%. According to Figure 11(c), during testing, it converged rapidly in 0-600 rounds. The loss value reaches its lowest point between 600 and 1200 rounds. According to Figure 11(d), the accuracy began to fluctuate after 600 rounds, with a maximum accuracy of 72.16% on the test set. Figure 12 shows the effect of full connection network, 1D convolution network and 1D convolution+Long short-term memory network on PR.

According to Figure 12(a), the recognition accuracy of fully connected, 1D convolution, and 1D convolution+LSTM algorithms increases in sequence. However, its highest accuracy rate only reached 85.67%, which did not achieve the desired effect. According to Figure 12(b), the recognition rate of 1D convolution+LSTM for single poses 5, 6, and 17 is as high as 100%. However, the recognition rate for poses 2, 4, 7, 8, 9, 10, 11, 12, 14, and 15 is relatively low, with a range of 50% -80%, indicating a high similarity among these groups of poses. TEA decoder-based PRA constructed through research preserves the best performing weights during the training process. Firstly, the convolutional output of the first layer and the convolu-

Figure 12
Effect of three kinds of networks on gesture recognition



tional output of the fourth layer were decoded, and an average of 8 and 16 frames of action sequences were selected as inputs to the model. Figure 13 shows the trend of loss values and accuracy on the training set and testing when inputting 8 frames.

According to Figure 13(a), the overall loss value on training set shows a decreasing trend, with a relatively rapid convergence in 0-40 rounds and a slower trend afterwards. According to Figure 13(b), the accuracy trend on training set is ultimately fixed at around 91.12%. According to Figure 13(c), the loss values on the test set show a gradually decreasing trend from 0 to 40 rounds. According to Figure 13(d), the accuracy shows an upward trend, with the best PR accuracy of 83.74%. Figure 14 shows the trend of loss values and accuracy in indifferent datasets when 16 frames are input.

Figure 13
Trend of loss value and accuracy on training set and test set when 8 frames were input

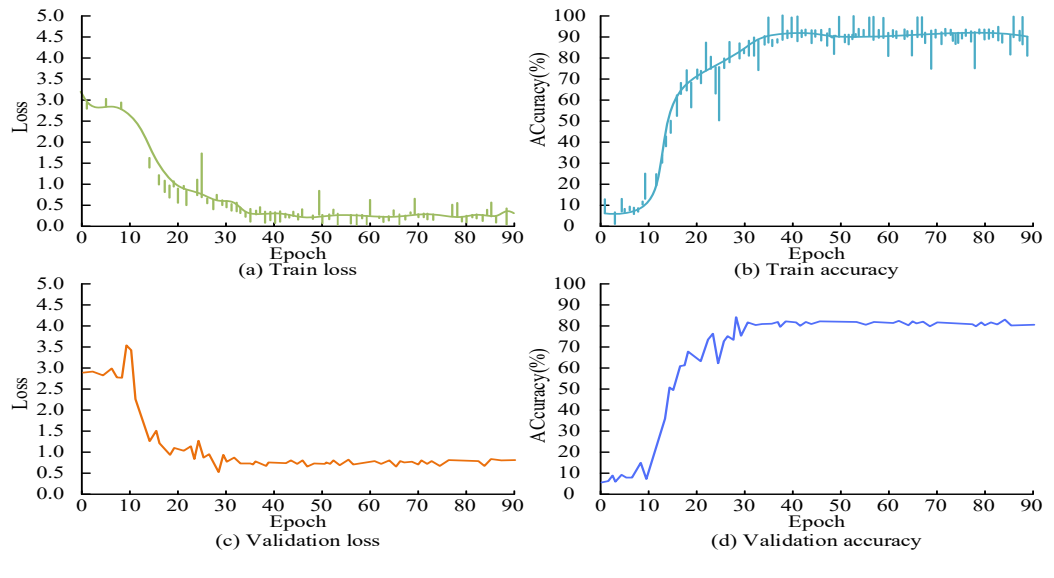
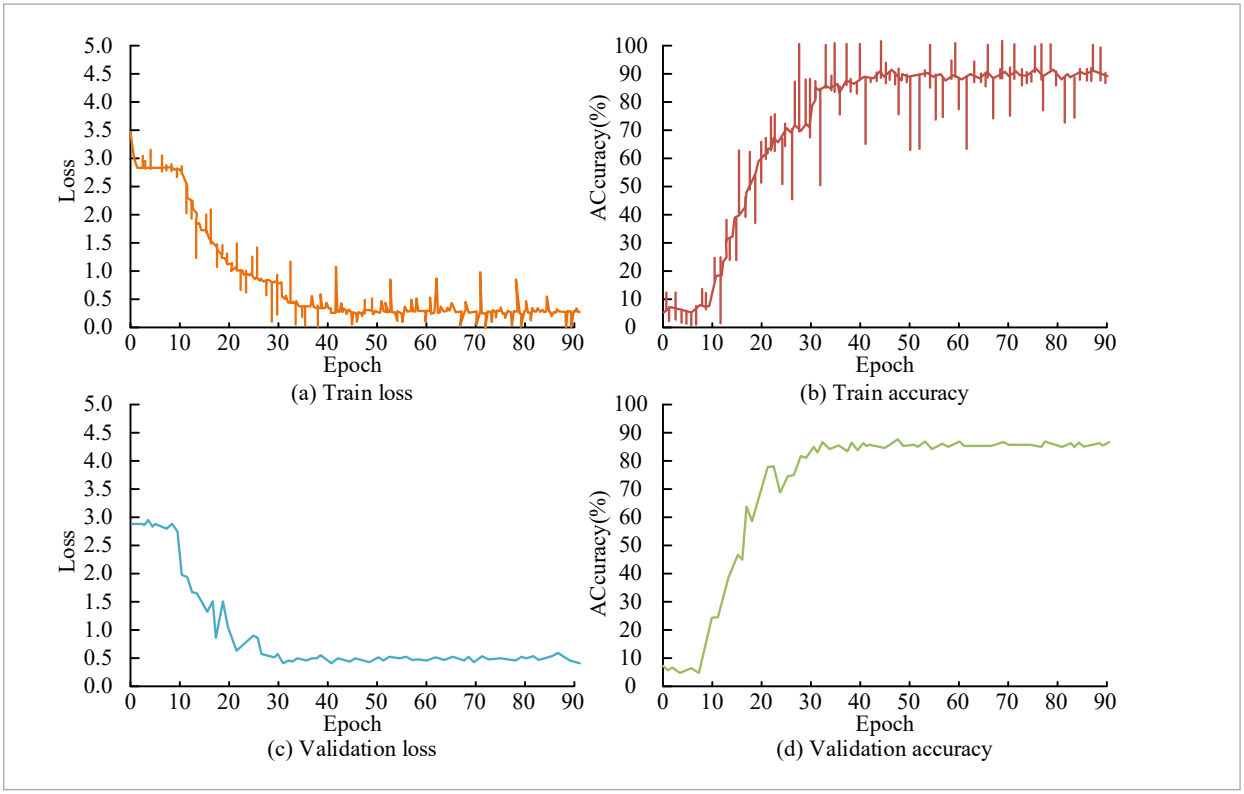


Figure 14
Trend of loss value and accuracy on training set and test set when 16 frames were input



According to Figure 14(a), the overall loss value on the training set shows a decreasing trend, with a relatively rapid convergence in 0-40 rounds and a slower trend afterwards. According to Figure 14(b), the final accuracy is around 91.12%. According to Figure 14(c), test set's loss values show a gradually decreasing trend from 0 to 40 rounds. According to Figure 14(d), test set's accuracy shows an upward trend, with an optimal PR accuracy of 88.35%. The study constructed PRAs for ResNet50, ResNet50 based on TEA module, and ResNet50 based on TEA Decoder, respectively. In Figure 15, action sequences of 8 and 16 frames were selected as inputs to the model, and a total of 10 experiments with the same network training strategy were designed.

According to Figure 15, the best results based on action sequences of 8 and 16 frames as model inputs were obtained by ResNet50 TEA Decoder (layer1-2-4). After adding Decoder module to the network, its optimal accuracy can reach 86.973% when 8 frame image sequences are used as inputs, and 92.052% when 16 frame image sequences are used as inputs. The results are better than those of the network without decoder module. Decoder module has a certain improvement effect on accuracy. Figure 16 shows the robustness test results of GRMA energy cost minimization model based on TDPRA.

According to Figure 16, research algorithm has better performance than traditional algorithms. Research algorithm can save up to 10% of total energy cost while minimizing total temperature deviation from the added value. Research algorithm can provide a more effective and flexible compromise between maintaining TC performance and reducing energy costs. To further verify the performance of the research method, the method is compared with the current building energy saving technology in terms of energy saving efficiency, intelligence degree, technical realization difficulty and cost. Data on energy efficiency, intelligence, technical difficulty and cost of thermal uncomfortable posture recognition algorithm and current building energy saving technology were collected. De-

Figure 15
The frame numbers of the five models correspond to the accuracy results

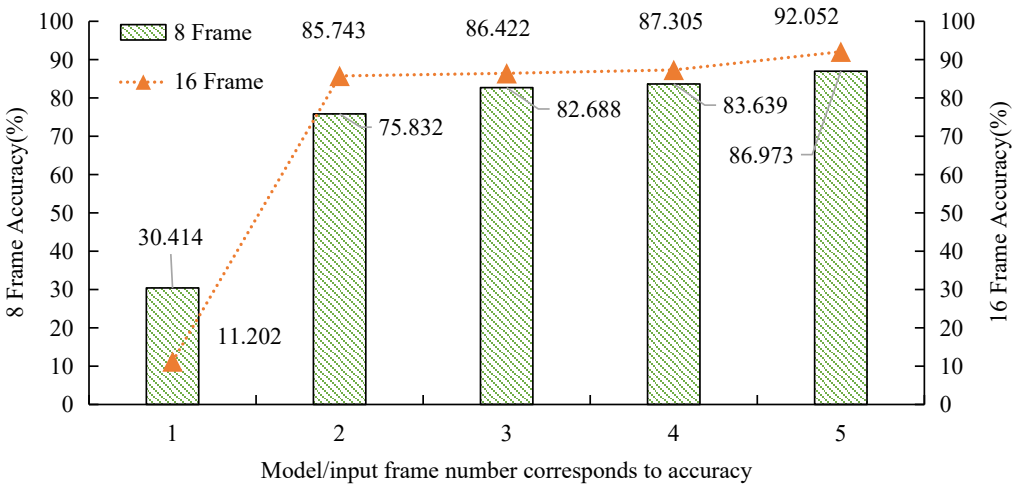
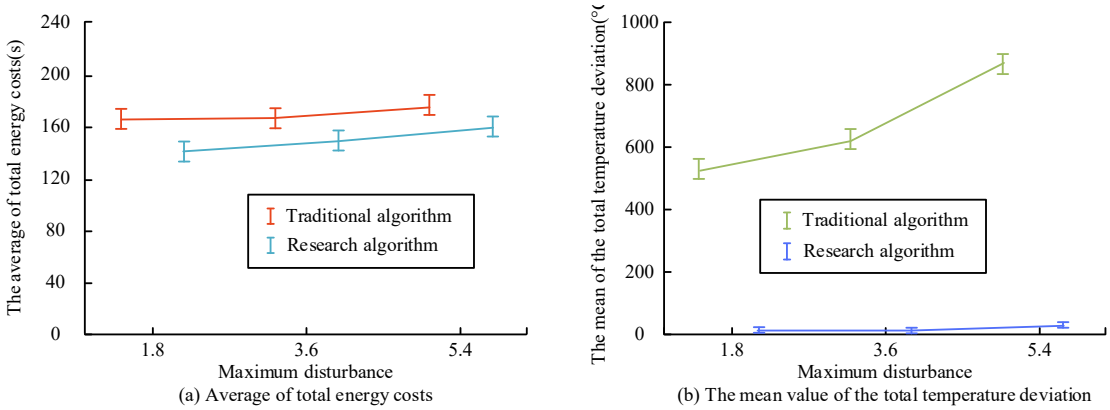


Figure 16
Algorithm robustness test



scriptive statistical analysis was performed to calculate mean ± standard deviation. Independent sample t test was used to evaluate significance, and $P<0.05$ indicated significant difference. The performance comparison results of various building energy saving methods are shown in Table 2.

As can be seen from Table 2, the values of the research method in terms of energy saving efficiency, intelligence degree, technical implementation ease and attitude recognition effect are all above 92%, which are superior to other existing building technical methods, and the P-values are all less than 0.05, indicating that the experimental results are statistically significant. In the total energy cost saving rate, the research method is significantly higher than other methods, and its energy cost saving rate can be as high as 10%.

Table 2
Performance comparison results of various building energy saving methods

Method	Energy efficiency (%)	Degree of intelligence	Technical implementation ease (%)	Attitude recognition effect (%)	Total energy cost savings (%)	*P*
Research method	98.65±0.23	92.49±1.30	93.64±0.74	97.64±0.63	10.11±0.02	<0.05
Literature [25]	91.30±0.42	89.14±1.29	88.24±0.84	89.12±0.37	6.23±0.11	<0.05
Literature [14]	86.04±0.31	82.76±1.03	87.12±0.67	83.41±0.86	5.64±0.34	<0.05
References [5]	88.22±0.42	86.19±1.14	90.34±0.53	91.05±0.65	7.21±0.32	<0.05
References [16]	89.24±0.36	90.14±1.21	91.34±0.32	92.71±0.33	8.33±0.42	<0.05
References [15]	95.73±0.45	91.02±1.39	92.63±0.49	93.62±0.79	9.31±0.21	<0.05
P	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

5. Conclusion

At present, TC detection mainly adopts contact and semi contact detection, both of which have issues such as equipment specificity and intrusion into the detected person. To avoid these issues, an innovative camera based non-contact measurement method is proposed in the study. Firstly, a framework for TDPRA was designed based on human BK. On this basis, in order to improve its detection accuracy, an end-to-end human TDPRA was proposed. Finally, based on TDPRA, a rural household HVAC system model was constructed. Experiments have shown that the recognition rate of 1D convolution+LSTM for single poses 5, 6, and 17 is as high as 100%. When inputting 8 frames, the convergence is relatively fast in 0-40 rounds on the training set, and then tends to slow down. The accuracy trend on the training set is ultimately fixed at around 91.12%. On the test set, the loss value shows a gradual decrease trend from 0 to 40 rounds, and the best PR accuracy is 83.74%. When inputting 16 frames, the convergence is relatively fast on the training set in 0-40 rounds, and its accuracy trend finally freezes at around 91.12%. The accuracy on the test set shows an upward trend, with the optimal PR accuracy of 88.35%. The best results based on action sequences of 8 and 16 frames as model inputs were obtained by ResNet 50-TEA Decoder (layer1-2-4). After adding Decoder module to the network, the optimal accuracy of 16 frame image sequences as input can reach 92.052%. Decoder module has a certain improvement effect on accuracy. Compared to traditional algorithms, research algorithm can save up to 10% of total energy costs while minimizing the total temperature deviation from the added value. Although the research has been successful, there are still limitations. For example, the same pose may occur for different reasons (such as fatigue or restlessness), adding complexity to the algorithm. Variables in the laboratory environment and real-world applications (e.g., interior design, climate conditions) can affect algorithm performance. Future research should develop personalized thermal comfort models, expand the data set to cover more practical scenarios, and test and optimize in practical applications to improve the reliability and applicability of the algorithm.

References

1. Akinbobola, A., Fafure, T. Assessing the Impact of Urbanization on Outdoor Thermal Comfort in Selected Local Government Areas in Ogun State, Nigeria. Nigerian Journal of Environmental Sciences and Technology, 2021, 5(1), 120-139. https://doi.org/10.36263/nijest.2021.01.0243
2. Amor, B. B., Arguillère, S., Shao, L. ResNet-LDDMM: Advancing the LDDMM Framework Using Deep Residual Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(3), 3707-3720. https://doi.org/10.1109/TPAMI.2022.3174908

3. Atsushi, Y., Patrick, A. Inter-Temporal Aggregation for Spatially Explicit Optimal Harvest Scheduling Under Area Restrictions. Forest Science, 2021, 5, 587-606. https://doi.org/10.1093/forsci/fxab025

4. Austin, M. C., Bruneau, D., Nadeau, J. P., Jaupard, D. Operation Assessment of an Air-PCM Unit for Summer Thermal Comfort in a Naturally Ventilated Building. Architectural Science Review, 2021, 64(1-2), 37-46. https://doi.org/10.1080/00038628.2020.1794782

5. Bia, A., Koltuk, B. Thermal Comfort Measurements in the Energis Building. Structure and Environment, 2021, 13(1), 10-15. https://doi.org/10.30540/sae-2021-002

6. Bogdanov, S. A., Sidelnikov, O. S., Redyuk, A. A. Application of Complex Fully Connected Neural Networks to Compensate for Nonlinearity in Fibre-Optic Communication Lines with Polarisation Division Multiplexing. Quantum Electronics, 2021, 51(12), 1076-1080. https://doi.org/10.1070/QEL17656

7. Chen, C. F. R., Fan, Q., Panda, R. Crossvit: Cross-Attention Multi-Scale Vision Transformer for Image Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357-366. https://doi.org/10.1109/ICCV48922.2021.00041

8. Han, X., Zhang, C., Tang, Y., Ye, Y. Physical-Data Fusion Modeling Method for Energy Consumption Analysis of Smart Building. Journal of Modern Power Systems and Clean Energy, 2022, 10(2), 482-491. https://doi.org/10.35833/MPCE.2021.000050

9. Ismail, N., Ouahrani, D. Modelling of Cooling Radiant Cubicle for an Office Room to Test Cooling Performance, Thermal Comfort and Energy Savings in Hot Climates. Energy, 2022, 244(1), 120-139. https://doi.org/10.1016/j.energy.2022.123185

10. Jaffal, I. Physics-informed machine learning for metamodeling thermal comfort in non-air-conditioned buildings. Build. Simul. 16, 299-316 (2023). https://doi.org/10.1007/s12273-022-0931-y

11. Kadry, S., Tummala, S., Bukhari, S. A. C., Rauf, H. T. Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling. Current Oncology, 2022, 29(10), 7498-7511. https://doi.org/10.3390/curroncol29100590

12. Kim, Y. Fall Detection Technology Based on Alphapose Bone Key Points and GRU Neural Network. Computer Science and Application, 2021, 11(4), 840-848. https://doi.org/10.12677/CSA.2021.114086

13. Kong, X., Chang, Y., Li, N., Li, H., Li, W. Comparison Study of Thermal Comfort and Energy Saving Under Eight Different Ventilation Modes for Space Heating. Building Simulation, 2022, 15(7), 1323-1337. https://doi.org/10.1007/s12273-021-0814-7

14. Kwong, Q. J., Yang, J. Y., Ling, O. H. L., Edwards, R., Abdullah, J. Thermal Comfort Prediction of Air-Conditioned and Passively Cooled Engineering Testing Centres in a Higher Educational Institution Using CFD. Smart and Sustainable Built Environment, 2021, 10(1), 18-36. https://doi.org/10.1108/SASBE-08-2019-0115

15. Liu, S., Zhu, C. Jamming Recognition Based on Feature Fusion and Convolutional Neural Network. Journal of Beijing Institute of Technology, 2022, 31(2), 169-177. DOI: 10.15918/j.jbit1004-0579.2021.105.

16. Meng, L., Li, H., Chen, B. C., Lan, S., Wu, Z., Jiang, Y. G., Lim, S. N. Adavit: Adaptive Vision Transformers for Efficient Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12309-12318. https://doi.org/10.1109/CVPR52688.2022.01199

17. Qabbal, L., Younsi, Z., Naji, H. An Indoor Air Quality and Thermal Comfort Appraisal in a Retrofitted University Building Via Low-Cost Smart Sensor. Indoor and Built Environment, 2002, 31(3), 586-606. https://doi.org/10.1177/1420326X211015717

18. Ramsey, D., Bouscayrol, A., Boulon, L., Desreveaux, A., Vaudrey, A. Flexible Simulation of an Electric Vehicle to Estimate the Impact of Thermal Comfort on the Energy Consumption. IEEE Transactions on Transportation Electrification, 2022, 8(2), 2288-2298. https://doi.org/10.1109/TTE.2022.3144526

19. Sun, S., Liu, R., Sun, S., Park, U. Keypoint-Based Disentangled Pose Network for Category-Level 6-D Object Pose Tracking. IEEE Computer Graphics and Applications, 2021, 42(5), 28-36. https://doi.org/10.1109/MCG.2021.3114181

20. Tummala, S., Kadry, S., Bukhari, S. A. C., & Rauf, H. T. Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling. Current Oncology, 2022, 29(10), 7498-7511. https://doi.org/10.3390/curroncol29100590

21. Villacis, A. H., Alwang, J. R., Barrera, V. Linking Risk Preferences and Risk Perceptions of Climate Change: A Prospect Theory Approach. Agricultural Economics, 2021, 52(5), 863-877. https://doi.org/10.1111/agec.12659

22. Wang, X., Cheng, M., Eaton, J., Hsieh, C. J., S. F. Fake Node Attacks on Graph Convolutional Networks. Journal of Computational and Cognitive Engineering, 2022, 1(4), 165-173. https://doi.org/10.47852/bonviewJCCE2202321

23. Xu, X., Zhu, L., Zhuang, W., Zhang, D., Lu, L., Yuan, P. Optimization of Optical Convolution Kernel of Optoelectronic Hybrid Convolution Neural Network. Optoelectronic Letters: English, 2022, 18(3), 181-186. https://doi.org/10.1007/s11801-022-1183-x

24. Yang, G. Adoption of Energy-Saving Materials in the Design of Hotel Intelligent System Under Low Carbon Environment. Integrated Ferroelectrics, 2021, 215(1), 256-266. https://doi.org/10.1080/10584587.2021.1911247

25. Younsi, Z., Qabbal, L., Naji, H. An Indoor Air Quality and Thermal Comfort Appraisal in a Retrofitted University Building via Low-Cost Smart Sensor. Indoor and Built Environment, 2022, 31(3), 586-606. https://doi.org/10.1177/1420326X211015717

26. Zheng, Z., Li, X., Sun, Z., Song, X. A novel visual measurement framework for land vehicle positioning based on multimodule cascaded deep neural network. IEEE Transactions on Industrial Informatics, 2020, 17(4), 2347-2356. https://doi.org/10.1109/TII.2020.2998107

27. Zhou, Q., Hui, T., Wang, R., Hu, H., Liu, S. Attentive Excitation and Aggregation for Bilingual Referring Image Segmentation. ACM Transactions on Intelligent Systems and Technology, 2021, 12(2), 1-17. https://doi.org/10.1145/3446345

28. Zhou, Z., Huang, H., Fang, B. Application of Weighted Cross-Entropy Loss Function in Intrusion Detection. Journal of Computer and Communications, 2021, 9(11), 1-21. https://doi.org/10.4236/jcc.2021.911001