IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 April 2023, accepted 30 May 2023, date of publication 6 June 2023, date of current version 12 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3283258

METHODS

Pseudo 3D Pose Recognition Network

YUANFENG XIE[1], XIANGYANG YU[1], WEIBIN HONG[2], ZHAOLONG XIN[3], AND YANWEN CHEN[4]
[1]State Key Laboratory of Optoelectronic Materials and Technologies, Department of Physics, Sun Yat-sen University, Guangzhou 510275, China
[2]Guangzhou Guangxin Technology Company Ltd., Guangzhou 510399, China
[3]Guangdong Topstrong Living Innovation and Integration Company Ltd., Zhongshan 528425, China
[4]Guangzhou Gaoke Communications Technology Company Ltd., Guangzhou 510000, China

Corresponding author: Xiangyang Yu (cesyxy@mail.sysu.edu.cn)

This work was supported in part by the research project titled ''Research on Intelligent Analysis, Judgment, and Recognition of Smart Regulation Behavior,'' under Grant 74130-71010019, and in part by the research project titled ''Research on Intelligent Detection Method of Hardware Surface Defects,'' under Grant 74130-71010020, both funded by Sun Yat-sen University.

ABSTRACT Multi-view human pose recognition has been extensively studied in computer vision due to its significant practical implications. Nonetheless, it remains a challenging task to effectively integrate distinctive view-based features and perform thorough qualitative analysis and quantitative evaluations. In this paper, based on an innovative multi-view fusion module and a novel Mutable Scaling Shortcut Connection, a pseudo 3D pose recognition neural network was meticulously crafted. The proposed network framework comprises four modules: Front Residual Module, 3D Convolution Cross View Fusion Module, Rear Residual Module, and Detection Module. The Front Residual Module serves as the head module with incipient pose heatmaps extraction functionality, taking preprocessed images of various views as separate inputs. The 3D Convolution Cross View Fusion Module performs 3D convolution fusion for the heatmaps output from Front Residual Module of each view, enabling the heatmaps to benefit from each other consequently. The Rear Residual Module extracts deeper-level features, and ultimately the Detection Module performs pose classification and recognition. The proposed network can be trained end-to-end and was evaluated with a Self-Built Multi-View pose recognition dataset. Analytical and evaluation approaches were used to explain the contributory effects of the 3D Convolution Cross View Fusion Module, which significantly improve recognition accuracy from approximately 70% to 91%-94% through Feature Aggregation, Strong Interaction Property among views, Sparsity Reduction, and Increasing Euclidean Distance.

INDEX TERMS Convolutional neural networks, Euclidean distance, image recognition.

I. INTRODUCTION

Human pose detection has been a popular research topic in the fields of computer vision and computer graphics for decades [1]. It allows for the recognition of specific body postures, such as sitting, meditation, standing, and squats, which can have numerous practical uses. The classification and recognition of human poses have the potential for extensive applications in various fields. Developers can leverage these applications, which have broad market prospects, for certain scenarios that require posture recognition and triggering, such as Behavior Recognition, Human-Computer Interaction, Video Games, Computer Animation, Virtual Reality, Rehabilitation Detection, and Robot Technology.

With the introduction and progress of deep neural networks and computer vision technology, researchers have made significant breakthroughs in human pose recognition research by adopting diversified specific solutions [2], [3], [4], [5], [6]. For example, the Cross View Fusion open-source model developed by Microsoft Research Asia [6] for 3D Human Pose Estimation has reduced estimation error to a great extent. However, multi-view human pose recognition remains a challenging task, requiring the fusion of multi-view images and reasonable qualitative analysis and quantitative evaluation.

This paper presents a meticulous crafting of a deep learning fusion neural network for Multi-View Pseudo 3D Pose Recognition. The proposed model not only fuses feature from multiple view images, but also facilitates improved accuracy, resulting in promising performance. Our work is specifically

The associate editor coordinating the review of this manuscript and approving it for publication was Antonio J. R. Neves.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

designed to address the challenges in this area, and our main contributions are outlined below.

- The network employs a CNN-based method to extract preliminary features and generate initial heatmaps from input images taken from diverse views To further enhance the quality of 2D heatmaps, inspired by literature [6], we adopted the Cross View Fusion method in the design of 3D Convolution Cross View Fusion Module (3D CCVFM) This method cross and adds heatmaps among each pair of views to fuse features from other views and improve the heatmap quality of any given view. Additionally, View Dimension is appended to the feature map, and 3D convolution operation is used to aggregate overall heatmaps. Experimental verification on a self-built dataset shows that the proposed 3D CCVFM can significantly improve model accuracy from approximately 70% to 91%94%, demonstrating promising performance on pose recognition tasks.
- Inspired by ResNet-50 [7] and LGAttNet [8], we adopted a common deployment order of Convolution, Rectified Linear Unit (ReLU), and Batch Normalization as the primary architecture for designing Front Residual Module (FRM) and Rear Residual Module (RRM). We also introduced the Pourer Layer into these primary frameworks. Unlike the identity shortcut connection, the proposed Pourer Layer features Mutable Scaling Shortcut Connection (MSSC). This novel connection mode allows the modulation factors to update themselves during the backpropagation process. With the network iteration model, the MSSC eventually finds the optimum weighted proportions for both the shortcut connection part and the residual function part. This unique feature of the Pourer Layer results in improved performance compared with the traditional identity shortcut connection.
- In Section IV-E, we conducted a detailed analysis and evaluation of 3D CCVFM from multiple perspectives. As demonstrated in the experimental results of Section IV-D, 3D CCVFM significantly improves the accuracy of model. Therefore, in Section IV-E.1, we conducted a qualitative analysis of the effective principle of 3D CCVFM from two perspectives: Feature Aggregation and Strong Interaction Property among views. Following this, in Section IV-E.2, we provided quantitative evaluation in terms of Sparsity and Euclidean distance. The experimental statistical results show that 3D CCVFM reduces the sparsity of feature maps and increases the Euclidean distance by an order of magnitude, indicating its superior performance.

II. RELATED RESEARCH

A. HUMAN POSE RECOGNITION

Multi-View 3D human pose recognition has been extensively studied in computer vision due to the significant amount of information that can be derived from the human body posture, which plays a critical role in human communication. However, much of the existing research focuses on 2D images, videos, and multi-view videos. For instance, in the work of Pehlivan and Duygulu [9], multi-view action videos were captured using 5 cameras, and these videos were used to construct volumes. Notably, these deep learning networks that achieve high performance require large datasets to provide robust support for the model. For example, the methods proposed by Jammalamadaka et al. [10] were evaluated quantitatively on a dataset consisting of a large number of images from standard benchmarks and frames from Hollywood movies. Similarly, the method proposed in [11] was applied to a dataset of 18 films, which contained more than three million frames. Wang et al. [12] created a large 3D Human Pose Recognition Dataset (HPRD) to evaluate pose classification and retrieval. This dataset included 1000 subjects, with each subject performing 100 poses.

However, the performance evaluation of our model was done on a self-built, small-scale Multi-View Human Pose Dataset, which means that we had limited data resources for pose classification and recognition. Additionally, our work differs from Human Pose Estimation (HPE) methods, such as those discussed in the literature [2], [3], [4], [5], [6]. For example, Qiu et al. [6] introduced a cross-view fusion method in CNN to jointly estimate 2D poses in multiple views and proposed a recursive image structure model to recover 3D poses from 2D poses. Chen et al. [2] estimated 3D multiple people poses from multiple calibrated camera views, taking the 2D poses in different camera coordinates as input, with the aim of acquiring exact 3D poses in global coordinates. Dong et al. [3] utilized a multi-way matching algorithm to cluster the detected 2D poses in all views. Zhang et al. [4] presented a geometric approach to reinforce the visual features of each pair of joints based on the IMUs and then lift the multi-view 2D poses to the 3D space by using an Orientation Regularized Pictorial Structure Model (ORPSM). Pavlakos et al. [5] introduced a geometry-driven approach to automatically collect annotations for human pose prediction tasks. In contrast to these previous works, our study focuses on the accuracy of the classification and recognition tasks in terms of model performance evaluation indicators.

B. MULTI-VIEW FEATURE FUSION

In the field of computer vision, the fusion of multi-view features has become a common practice for obtaining more discriminative features than the original input. The fundamental concept of feature fusion is to combine the features extracted from multiple images to better utilize the advantages of multi-view features. To fully exploit the potential of these diverse features, it is essential to jointly model them. Appropriate feature fusion methods have been shown to significantly enhance the efficiency and effectiveness of various computer vision tasks. Currently, there exist numerous feature fusion methods utilized in computer vision, among which are the Pooling Method, Ordered View Feature Fusion (OVFF), and Image Addition.

In order to synthesize feature information from multi-view 2D images, Su et al. [13] introduced the MVCNN architecture, which utilizes an innovative approach to perform Max Aggregation of views using a view-pooling layer. The MVCNN combines information from multiple views into a single and compact shape descriptor, exhibiting commendable recognition performance. However, max-pooling selectively captures the highest value in each feature map, which may result in the loss of spatial information. Additionally, without distinguishing whether a feature appears once or multiple times, the intensity information of other features may also be lost [14]. Furthermore, diverse pooling techniques have been proposed by researchers for feature fusion, including Within-Cluster Pooling [15], Soft-View Pooling [16], Intra-Group View Pooling [17], Harmonized Bilinear Pooling [18], Max-Pooling [19], Spatial Pyramid Pooling [20]. The OVFF [12] organizes the feature data into blocks based on viewing sequence and subsequently classifies it through a full connection layer. Similarly, [21] also employs a comparable connection method, whereby the resulting fused feature map is fed into a CNN layer and a Flatten layer for aggregation. The Image Addition method achieves feature fusion by directly adding heatmaps, as demonstrated by Cross View Fusion [6], Orientation Regularized Network (ORN) [4], and other similar approaches.

Our proposed 3D CCVFM distinguishes itself from prior research by emphasizing the importance of feature interaction among views. By expanding the view dimension of the feature graph, our method enables multi-view features to effectively and comprehensively interact during the 3D convolution operation. Notably, to the best of our knowledge, no prior work has leveraged 3D convolution to integrate multi-view features and achieve superior network performance. This is largely due to the challenging task of aggregating corresponding features from diverse views and conducting qualitative analysis and quantitative evaluations, which constitute pivotal contributions of our research.

C. SHORTCUT CONNECTION

In statistics, the residual is initially defined as the discrepancy between an actual observed value and its corresponding estimated value (also known as the fitted value). Residual networks, or ResNets [7], have gained popularity in the deep learning field due to their ability to break permutation symmetry [22], enhance generalization [23], and outperform other networks in the ImageNet image recognition competition. In fact, ResNet has become a fundamental network in the field of deep learning. ResNet_V2 [24], on the other hand, is a variation of ResNet that restructures the integral components sequence based on the ResNet architecture. Residual networks have been demonstrated to simplify the learning process and enhance effectiveness in various studies by leveraging the learning of the difference value in the signal rather than the original signal itself.

In recent years, ResNet-based research has yielded numerous outstanding achievements [23], [25], [26], [27], [28], [29], most of which are variant networks derived from ResNet. For example, Xie et al. constructed ResNext [25], which combines the ResNet and Inception Network (GoogleNet) [30] inspirations to acclimatize updated datasets or tasks. Huang et al. proposed a new architecture, DenseNet [26], that concatenates all layers directly through shortcut connection. In addition, by capitalizing on the advantages of the Dropout technique [31], Huang et al. [27] extended the method of randomly discarding certain hidden units in the fully connected layer to residual blocks. Chen et al. treated ResNet and DenseNet as two channels in parallel, and then fused the information to obtain Dual Path Network [28]. Veit et al. [23] successfully removed certain trained layers in ResNet and discovered that the network still maintained relatively fantastic performance. Furthermore, complex variant networks known as ResNet in ResNet (RIR) structures [29] have also been proposed.

However, these existing works primarily focuses on the architecture refinement of the original Residual Unit, which employed a simple matrix addition operation to connect the shortcut and residual function components, commonly known as the conventional identity shortcut connection [7], [24]. As a result, attention was not given to the feature vector channels screening. Despite the existence of ResNet variants with attention mechanisms, such as Selective Kernel Networks(SKNet) [32], Split-Attention Networks(ResNeSt) [33], Squeeze-and-Excitation Network [34], Deep Residual Shrinkage Networks [35], and other notable research endeavors, their focus was mainly on refining the residual function and not on weight allocation within the identity shortcut connection. When it comes to strengthening or weakening specific feature channels for screening purposes, these approaches have paid little attention to weight allocation within the identity shortcut connection.

This paper proposes a novel approach that focuses on configuring the weight proportions between the shortcut connection and residual function components. Our method, the Pourer Layer, leverages the Mutable Scaling Shortcut Connection (MSSC) mechanism, which enables simultaneous modulation of channel-weighted proportions and can be optimized during network training iterations. However, the hardness of this method is incapable of invoking existing functions directly such as convolutional layer, pooling layer, and fully connected layer, for automatic weight parameter updates. Instead, this approach poses a challenge as it requires the development of underlying algorithms.

III. NET DETECTION MODEL DESCRIPTION

As depicted in Fig. 1, the Pseudo 3D Pose Recognition Network is comprised of four sequential modules, with data input of each view corresponding to an autonomous Front Residual Module (FRM). The preprocessed images are methodically organized into data clusters, where multiple views of the same pose form an image data group. It is noteworthy that during the end-to-end training and inference of network, the input unit is no longer limited to a single two-dimensional pose detection image, but rather comprises an entire image data group, which is then associated with a single label.

The 3D Convolution Cross View Fusion Module (3D CCVFM) commences by performing a cross-image addition operation on the output features generated by FRM of each view. Subsequently, the feature image data are passed through the first RS layer, where it undergoes deformation, rearrangement, and is merged into a block data structure that facilitates 3D convolution input. The 3D convolution layer within the module compresses the feature data of $\mathrm{N}_{\mathrm{view}}^{2}$ dimensions into a singular unit, thereby achieving views fusion. During the 3D convolution process, the module enables information interaction by considering view-versus-view and channel-versus-channel, thus ensuring comprehensive information exchange. To enable deeper feature extraction by the Rear Residual Module (RRM), the second RS layer deforms the output of the 3D convolution layer, thereby maintaining consistency with the output data size of FRM.

The RRM captures intricate features, which are subsequently classified and recognized by the Detection Module (DM). DM is a shallow neural network module that incorporates a Global Average Pooling (GAP) [36] layer and a fully connected layer, enabling classification of multi-view poses. The integration of GAP not only curtails the number of parameters but also mitigates the risk of overfitting that arises due to redundant spatial information. The ultimate output of network is represented through Softmax prediction labels. Following subsections explain these modules at great length.

A. NETWORK COMPONENTS

The 3D CCVFM encompasses several image aggregation functions, coupled with two RS layers that facilitate data transformation in diverse ways, and a fusion layer based on 3D Convolution operation. Notably, the FRM and RRM, two finely crafted 2D-CNN components of the network architecture, integrate Pourer Layers that exploit the advantages of Mutable Scaling Shortcut Connection mode, resulting in enhanced performance. Additionally, the DM component comprises a GAP layer, a 2048 fully connected layer, and a Softmax layer for accurate and efficient classification.

1) FRONT RESIDUAL MODULE (FRM)

In general, FRM is implemented utilizing a convolutional neural network without fully connected layer but Pourer Layer. The FRM architecture described in this paper comprises four 2D convolutional layers (trunk) and one down-sampling layer (branch) to extract heatmaps from input images. As demonstrated in Fig. 2, the trunk draws inspiration from ResNet-50 [7] and LGAttNet [8] and includes a max-pooling layer attached to the first convolutional layer. The subsequent layers employ a common design structure of Convolution, ReLU, and Batch Normalization. The Strided Pooling and Strided Convolutional Layers compress the spatial dimension, reducing the input data shape from H×W to H/2×W/2. In the branch, a down-sampling layer is attached to the Pourer Layer, establishing shortcut connections to prevent gradient exploding and vanishing. Shortcut connections can also alleviate network performance degradation resulting from deepening the network, a challenge that conventional normalized initialization and batch normalization are incapable to overcome [7]. It should be emphasized that FRM requires 3-channel input images of multi-view pose in 112 × 112 dimensions. The implementation of FRM is transformed as follows.

$$M_{\mathrm{FRM}\_i} = \mathrm{F}_{\mathrm{FRM}\_i}\left(I_i\right) \tag{1}$$

The variable I_i is the input of the FRM function, corresponding to the FRM of the i^{th} view. Here, the digit i denotes the position of the view, ranging from 1 to $\mathrm{N}_{\mathrm{view}}$, where $\mathrm{N}_{\mathrm{view}}$ represents the total number of views. After processing the input image, each FRM generates an initial pose heatmap represented by M_{FRM_i}.

2) 3D CONVOLUTION CROSS VIEW FUSION MODULE (3D CCVFM)

As depicted in Fig. 1, during each batch of the training or inference process, the 3D CCVFM receives a simultaneous input of $\mathrm{N}_{\mathrm{view}}$ feature maps for cross fusion processing during the forward propagation phase. To obtain new superimposed feature maps that benefit from additional information provided by other views, each output feature map from the FRM undergoes an image addition operation with the output of other FRMs. As a result of this operation, $\mathrm{N}_{\mathrm{view}}$ correlated superimposed feature maps are produced from each FRM, leading to a total of $\mathrm{N}_{\mathrm{view}}^{2}$ superimposed feature maps in each operation batch. This technique facilitates the exchange of feature information among different views and enhances their information sharing capability.

In the case of $\mathrm{N}_{\mathrm{view}}^{2}$ superimposed feature maps with a shape of H × W × C, they are intrinsically independent of each other and, as such, cannot be directly subjected to the 3D convolution kernel operation. To enable the fusion operation of 3D convolutional layers, the first RS layer introduces a novel dimension (the ''view'' dimension) and preserves the original data arrangement of dimensions H, W, and C. The $\mathrm{N}_{\mathrm{view}}^{2}$ feature maps are then reorganized along this view dimension. This reshaping of feature data effectively elevates the dimensionality from 3 to 4, thereby transforming the size from H × W × C to H × W × $\mathrm{N}_{\mathrm{view}}^{2}$ × C, which aligns precisely with the kernel size of the 3D convolution layer (i.e., $1 \times 1 \times \mathrm{N}_{\mathrm{view}}^{2} \times C$).

The 3D convolution kernel has a size of 1 × 1 in both the height and width dimensions. This indicates that the convolution operation preserves the original dimensions of its input maps and disregards information interaction within the same channel or view. It enhances performance by incorporating the ''view'' dimension. In fact, the 3D CCVFM facilitates information integration across different views and channels, making it a noteworthy advancement.

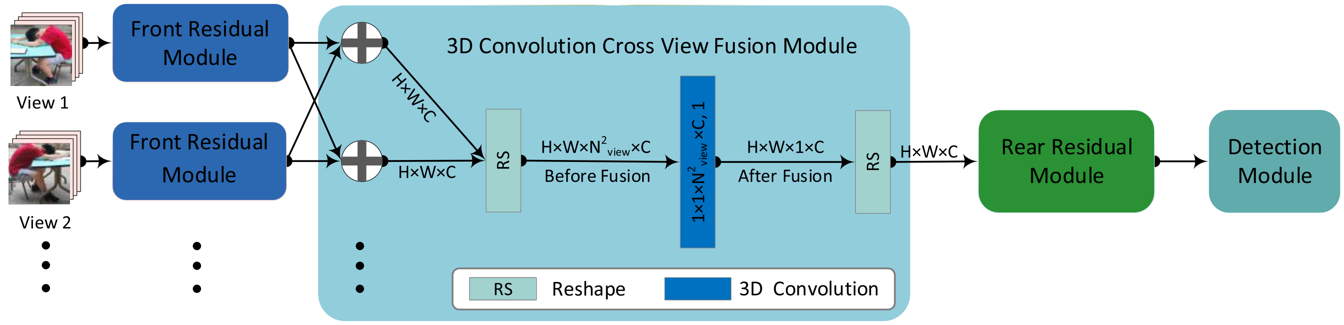

FIGURE 1. The overall architecture of the proposed Pseudo 3D Pose Recognition Network model. The input data of each view correspond to an independent Front Residual Module. In 3D Convolution Cross View Fusion Module (3D CCVFM), the annotations on arrows represent the shape of feature map, such as H × W × C represents Height × Width × Channel. In 3D convolution layer, the annotations denote parameters, including height, width, length, channel and stride.

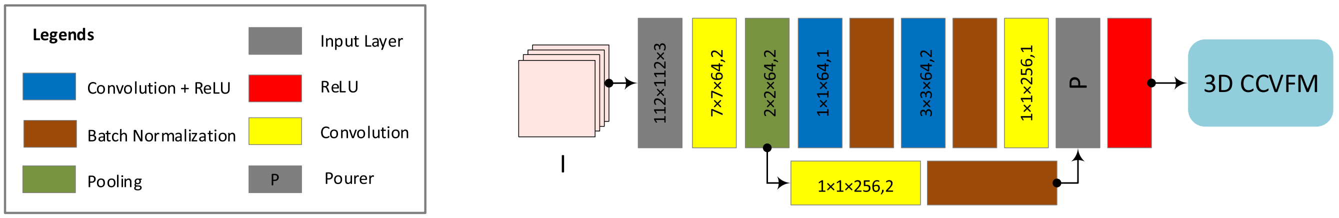

FIGURE 2. Front Residual Module: 'I' is the input image given to the Front Residual Module, a shallow five-layer CNN including the proposed Pourer Layer. The heatmap is output from the last ReLU layer and fed to the 3D CCVFM. In each operation layer, the annotations denote parameters, such as height, width, length, and stride.

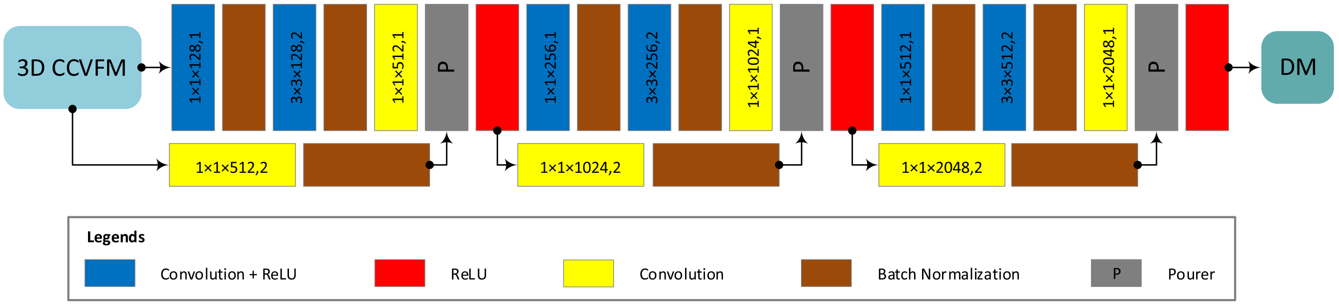

FIGURE 3. Rear Residual Module: Input given to this twelve-layer convolutional network including the proposed Pourer Layer is the feature map fused by 3D CCVFM. Rear Residual Module captures high-level features and the result is fed to the DM for pose classification. Within each operation layer, the annotations represent parameters such as height, width, length, and stride.

As a result of the fused output no longer necessitating the view dimension, the second RS layer extracts and discards this dimension, achieving consistency between the ultimate output form of 3D CCVFM and that of the single FRM. The reshape operation does not alter the total number of elements in the tensor. However, it performs essential dimension adjustments and order remodeling among the elements. These adjustments and remodeling are significant for the network to align the convolution kernel with data that have varying dimensions and scales.

The resultant feature map, M_{3D_CCVFM}, is given as follows.

$$M_{3D\_CCVFM} = \mathrm{F}_{3D\_CCVFM}(M_{FRM\_1}, M_{FRM\_2}, \ldots, M_{FRM\_Nview}) \quad (2)$$

N_{view} denotes the total number of views, and M_{FRM_1}, $M_{FRM\_2}, \ldots, M_{FRM\_Nview}$ respectively correspond to the outputs generated by FRM for each view. Each of these outputs independently serves as an input to the 3D CCVFM function F_{3D_CCVFM}, which subsequently produces M_{3D_CCVFM} through a transformation process. The latter is then exported to capture deeper features downstream in the network.

3) REAR RESIDUAL MODULE (RRM)

RRM employs widely-used architecture, including Convolution, ReLU, and Batch Normalization, as its primary design features, similar to FRM. To ensure data shape consistency with the trunk, the branch performs both down-sampling and increasing channel in the feature maps. In contrast to FRM, RRM utilizes a greater number of network layers, comprising of three bottleneck structures, as illustrated in Fig. 3. This increased depth allows RRM to extract deep features from the 3D CCVFM output, which are subsequently transmitted to the DM at the end of the network.

As the depth of network layers increases, the neural network can capture increasingly abstract features, leading to a greater potential for arranging and combining these features. To achieve strong expressive power, it is necessary to gradually decrease the height and width dimensions while increasing the number of channels in the network, as noted by [37]. To ensure that the feature information is adequately conveyed, the feature maps undergo three successive rounds of down-sampling from the input to the output. By setting the stride to 2, the height and width dimensions of the feature map are reduced by half at each down-sampling stage, while the number of channels is doubled. The resulting high-level feature map on the RRM output side is highly simplified in terms of height and width, but contains numerous channels, with a size of 2 × 2 × 2048. At this point, the feature map captures most of the crucial features for the network design task, rendering it suitable for classification and recognition.

The implementation of RRM is expressed as follows. Specifically, $\mathrm{F_{RRM}}$, M_{RRM}, and $M_{\mathrm{3D\_CCVFM}}$ respectively represent the function, output, and input of RRM. $M_{\mathrm{3D\_CCVFM}}$ also denotes the output generated by the 3D CCVFM.

$$M_{\mathrm{RRM}} = \mathrm{F_{RRM}}(M_{\mathrm{3D\_CCVFM}}) \tag{3}$$

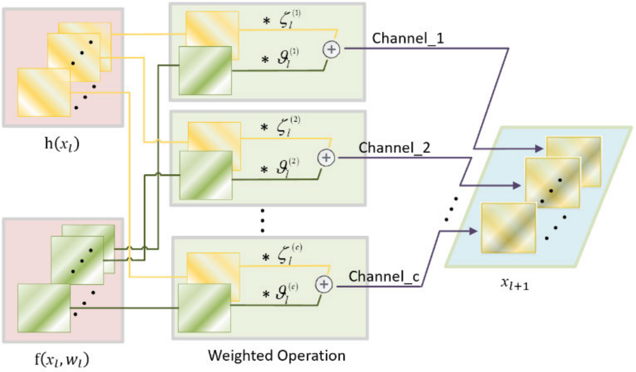

FIGURE 4. Pourer Layer: The output of shortcut connection part h(x_l) and the residual function part f(x_l, w_l) are fed into Pourer Layer. Then their corresponding channels are modulated and added to form the output x_{l+1}.

4) POURER LAYER (PL)

The utilization of residual learning in neural networks is a key technique that enables optimization through the inclusion of shortcut connections. A residual block, also known as a residual unit, is a module that consists of multiple layers with shortcut connection and preserves the integrity of information flow from the input to the output port. By training the residual feature, the learning objective and complexity are simplified, thereby facilitating network optimization.

The proposed Pourer Layer distinguishes itself from the traditional Identity Shortcut Connection (ISC) by capitalizing on the Mutable Scaling Shortcut Connection (MSSC). Fig. 4 illustrates how the Pourer Layer facilitates the configuration of weighted parameters ζ and ϑ for the shortcut connection section and residual function section, respectively. The corresponding channels are then added to form the output feature map.

Initialized with 1, ζ and ϑ are both parameter vectors whose dimensionality corresponds to the channel number of the Pourer Layer inputs. Continuous self-optimization through a one-hundred-fold learning rate in the backpropagation process of model, ζ and ϑ can be optimized to achieve the best possible weighted proportions. Through iterative training, shortcut connection part and the residual function part can acquire the optimal weighted proportions, i.e., the optimized solution of ζ and ϑ. The implementation of the Pourer Layer is transformed as

$$x_{l+1} = \zeta_l \cdot \mathrm{h}\left(x_l\right) + \vartheta_l \cdot \mathrm{f}\left(x_l, w_l\right), \tag{4}$$

where x_l and x_{l+1} respectively represent the input and output of the l^{th} residual unit, whereas h and f denote the shortcut connection function and residual function. Respectively ζ_l and ϑ_l are weighted parameters of h and f. w_l is a set of weights (and biases) associated with the l^{th} residual unit. By recursively applying this formulation, the equational expression is obtained as

$$x_{l+2} = \zeta_{l+1} \cdot \mathrm{h}\left(x_{l+1}\right) + \vartheta_{l+1} \cdot \mathrm{f}\left(x_{l+1}, w_{l+1}\right). \tag{5}$$

Equations (4) and (5) share similarities with the Constant Scaling Shortcut Connection (CSSC) and Convolutional Shortcut Connection (CSC), as discussed in the literature [24]. However, it is worth noting that the textual MSSC exhibits some notable characteristics. Firstly, the network utilizes MSSC only a few times, which mitigates the exponential increase or decrease of the modulation factors, ζ and ϑ. This can prevent optimization difficulties and network crashes, as observed with CSSC testing on ResNet-110 (which consists of 110 layers). Secondly, ζ and ϑ are independent of each other within any residual unit. Finally, in the process of network backpropagation, ζ and ϑ update themselves to meet the optimal weighted proportions required for network performance.

In the preceding analysis, the Pourer Layer does not hinder the propagation of information or impede the training procedure. On the contrary, the additive modulated factor enhances the effective transmission of information. The gradient equation for x_l during backpropagation from any deeper unit $l + 1$ to any shallower unit l is decomposed as follows. Based on the chain rule of backpropagation, (6) is derived, where the loss is denoted as E.

$$\frac{\partial E}{\partial x_l} = \frac{\partial E}{\partial x_{l+1}} \frac{\partial x_{l+1}}{\partial x_l} = \frac{\partial E}{\partial x_{l+1}} \left[\zeta_l \cdot \frac{\mathrm{h}\left(x_l\right)}{\partial x_l} + \vartheta_l \cdot \frac{\mathrm{f}\left(x_l, w_l\right)}{\partial x_l} \right] \tag{6}$$

In the same way, the gradient equation for ζ_l and ϑ_l can be decomposed as follows. The ratio can be obtained as $\operatorname{grad}(\zeta_l)/\operatorname{grad}(\vartheta_l) = \mathrm{h}(x_l)/\mathrm{f}(x_l, w_l)$ by comparing the two expressions.

$$\operatorname{grad}(\zeta_l) = \frac{\partial E}{\partial \zeta_l} = \frac{\partial E}{\partial x_{l+1}} \cdot \frac{\partial x_{l+1}}{\partial \zeta_l} = \frac{\partial E}{\partial x_{l+1}} \cdot \mathrm{h}(x_l) \tag{7}$$

$$\operatorname{grad}(\vartheta_l) = \frac{\partial E}{\partial \vartheta_l} = \frac{\partial E}{\partial x_{l+1}} \cdot \frac{\partial x_{l+1}}{\partial \vartheta_l} = \frac{\partial E}{\partial x_{l+1}} \cdot \mathrm{f}(x_l, w_l) \tag{8}$$

5) DETECTION MODULE (DM)

The DM comprises a Global Average Pooling (GAP) layer and a fully connected layer with a size of 2048. The GAP layer aggregates spatial information and is more resistant to spatial variations in input feature. As a result of the GAP layer, the feature map is transformed from 2 × 2 × 2048 to 1 × 1 × 2048, reducing the required fully connected layer size and effectively avoiding over-fitting caused by a large FC layer. Additionally, a softmax classification layer is appended to the DM to generate the initial output for the classification task, ultimately outputting predicted probability values. The pseudo 3D pose recognition task via the DM is explained in (9), where M_{RRM}, F_{DM}, and p represent the RRM output, DM function, and prediction probabilities associated with the input sample, respectively.

$$p = \mathrm{F}_{\mathrm{DM}}(M_{\mathrm{RRM}}) \tag{9}$$

B. LOSS FUNCTION

During the training process, the typical softmax cross-entropy loss was utilized as a classification loss to optimize the model. Here, the associated ground truth label vector is denoted by $\mathbf{y}$, and the predicted values for the corresponding image sample group are represented by $p = [p_1, \ldots p_K]$, where K is the number of pose categories.

$$E_{\mathrm{cls}} = -y \cdot \log(p) \tag{10}$$

IV. EXPERIMENTAL SETUP AND RESULTS

A. DATASETS ESTABLISHMENT

We constructed a small Multi-View Human Pose Dataset for the Pseudo 3D Pose Recognition task, comprising approximately 1.5 thousand multi-view human pose images from 4 participants. The details of the dataset collected for the experiments are provided below. As indicated in Fig. 5, the participant is positioned upright along a fixed axis (the z-axis), and eight cameras are directed toward the centre with a 45° angle between each adjacent view. These cameras surround the central participant for image capture and data collection. Eight images captured from the same pose comprise an image data group, which is preprocessed and utilized as a single input unit for the network. To accurately reflect model performance, the standard distributional percentage was utilized for small datasets to develop and test our ideas and models. Specifically, the ratio of the training set, validation set, and test set is approximately 6:2:2.

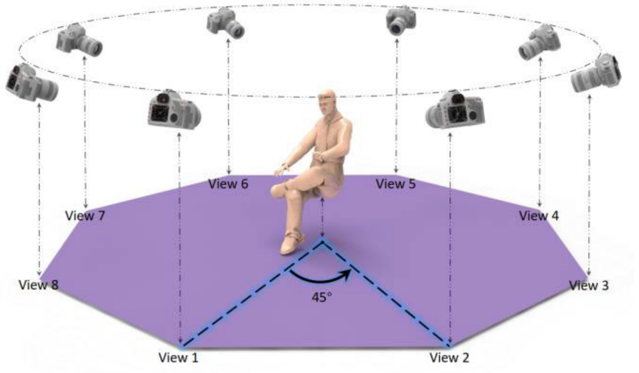

FIGURE 5. Illustration of Multi-View Human Pose Dataset establishment process. The cameras, directing to the centre participant, corresponding to the eight views, are all at the same level of height.

B. PRE-PROCESSING

The preprocessing of the image entails three essential stages: data augmentation, intensity normalization, and scale normalization. During the data augmentation phase, the Hard/Local Attention Mechanism enhancement method is utilized to eliminate irrelevant areas of the image directly. The modulation weight is assigned as zero for background areas that are uncorrelated with the pose and as one for postural-related regions. This process discards background details and reduces noise disturbances, as these regions contain no valuable information regarding pose recognition. The superiority of this approach lies in its allocation of limited information processing resources to the critical portions of the image. As a result, the neural network can focus on a specific region and pay strong attention to it, thereby reducing the computational cost and improving the accuracy of pose recognition. During the intensity normalization stage, the original images are converted into corresponding standard forms, inspired by the approach in Convolutional 3D (C3D) [38]. Thirdly, in the scale normalization section, the images are reduced to a size of 112 × 112 pixels using the linear interpolation method. This technique simultaneously preserves the bulk principal information of the image data and compresses data size. The scale normalization process mitigates the computational load during both model training and inference.

C. EXPERIMENTAL SETUP AND PARAMETERS

All the experiments were conducted with Ubuntu 20.04, Python 3.8.11 and PyTorch 1.9.0 on an NVIDIA GeForce RTX 3090 GPU Server (24 GB). Prior to being fed into the inception network, the images were preprocessed and resized to 112 × 112 pixels. The network employed gradient descent method with an SGD optimizer for iterative optimization. An initial learning rate of 0.008 and a momentum of 0.9 were utilized to accelerate model convergence. Weighted decay implementation was performed by capitalizing on a scheduler to divide the learning rate by 10 every 10 epochs. Model training was carried out for 40 epochs, with early stopping employed.

TABLE 1. The pseudo 3D pose recognition outcomes using various performance metrics.

Metrics	3D CCVFM	×	×	√	√
	Pourer Layer	×	√	×	√
Class-1	Precision	0.699	0.618	**1.000**	**1.000**
	Recall	0.580	0.723	0.786	**0.857**
	F1-Score	0.634	0.667	0.880	**0.923**
Class-2	Precision	0.737	0.780	0.870	**0.909**
	Recall	0.825	0.688	**1.000**	**1.000**
	F1-Score	0.779	0.731	0.930	**0.952**
Mean	Precision	0.718	0.695	0.933	**0.953**
	Recall	0.692	0.705	0.886	**0.926**
	F1-Score	0.703	0.698	0.905	**0.938**
Accuracy		72.43%	70.22%	91.18%	**94.12%**

D. ABLATION STUDIES AND OUTCOMES

The present study draws upon data obtained from our Multi-View Human Pose Dataset, a collection of diminutive yet richly informative samples. To evaluate the efficacy of our experimental approach, we focused on two distinct pose classes from the test set and employed a range of performance metrics, including F1-score, recall, precision, and accuracy. By considering the latter, we were able to gain a more comprehensive understanding of the predictive capabilities of model and assess its overall effectiveness.

To evaluate the effectiveness of the model, the confusion matrix technique was applied, and second-level evaluation indicators, namely accuracy, recall, and precision, were derived from the basic statistical results of the confusion matrix. The three-level evaluation index, F1-score, was established based on recall and precision mapping. During the experiment, the weight data of the model that exhibited the highest accuracy on the validation set were selected for inference on the test set. An experimental ablation study method was implemented to explore the contribution of each factor to the overall model performance by eliminating various module components. The classification performance of the pose recognition task was quantified using the index values presented in Tab. 1.

The Multi-View Human Pose Dataset comprises two distinct classes, each representing a diverse pose. The Fusion Module Network (FMN), which incorporates a 3D CCVFM, utilizes a group of images from eight different angles featuring the same pose as its input unit. Conversely, the architecture of the Without Fusion Module Network (WFMN) is based on a sequence of FRM (single), RRM, and DM. As the WFMN only contains a single FRM, it is only capable of processing a single image as input. To ensure a fair comparison with the FMN, the aggregate amount of input data for the WFMN was kept consistent with that of the FMN during the training, validation, and testing phases.

Tab. 1 illustrates that the incorporation of cross-fused features across multiple views leads to a substantial enhancement in pose recognition performance by utilizing the 3D CCVFM. Specifically, the effectiveness of the 3D CCVFM is demonstrated by the considerable improvement in model accuracy, which elevates from approximately 70% to high levels ranging from 91% to 94%. Moreover, the incorporation of the Pourer Layer yields improvements in various evaluation metrics, including precision, recall, and F1-Score. These findings suggest that both the 3D CCVFM and Pourer Layer significantly contribute to the improved recognition accuracy of the network.

E. ANALYSIS AND EVALUATION

In Section IV-E.1, a qualitative analysis is provided to examine the effective principles of 3D CCVFM from the perspectives of Feature Aggregation and Strong Interaction Property among views. Subsequently, Section IV-E.2 presents a quantitative assessment using sparsity and Euclidean distance as metrics. These results demonstrate that our proposed 3D CCVFM architecture plays an important role in the performance improvement of multi-view classification network.

1) QUALITATIVE ANALYSIS

To gain a more profound insight into the neuronal activation and channel interactions in feature maps, it is imperative to visualize the input and output of the 3D CCVFM. As depicted in Fig. 6, the preprocessed images exhibited a notable suppression of irrelevant information, notably facial details, while diligently preserving the crucial postural features. The visualization was accomplished by displaying two kinds of heatmaps, one illustrating the transformation from FRM to 3D CCVFM and the other demonstrating the transformation from 3D CCVFM to RRM. These heatmaps were arranged along the channel dimension to facilitate visualization. Upon visualizing the heatmaps, it became apparent that the output from FRM was intricate, owing to the entanglement of dense feature information. By contrast, the output from 3D CCVFM exhibited a clear distinction among channels. This observation was further validated by experimental statistics, which showed a lower level of sparsity and larger Euclidean distance, as elucidated in Section IV-E.2. The qualitative analysis of the visualization experiment underscores the detailedness of the generated feature maps. Specifically, the 3D CCVFM architecture produced feature maps that were distinct, informative, and less sparse. These findings portend the potential of 3D CCVFM in enhancing the performance of neural networks. The detailedness of qualitative analysis is described as follows.

a: FEATURE AGGREGATION

In general, relevant features are defined as image information advantageous for the current pose recognition task, whereas irrelevant features do not contribute to the task. Operations in a convolutional neural network, such as convolution, pooling, and activation, can be thought of as a multi-stage process that distills information from feature maps. This process continuously filters out irrelevant features, retaining high-level relevant features for classification and detection. As presented in Fig. 6, the input heatmaps of 3D CCVFM contain a large number of irrelevant features, resulting in complicated feature maps that conceal critical information for pose recognition. This extensive presence of irrelevant features negatively impacts learning efficiency and network performance. Fortunately, the feature distinction among channels becomes clear after cross-view fusion with 3D CCVFM. This process strengthens and integrates the differentiated features learned by each view while suppressing irrelevant features, effectively reducing the burden on subsequent RRM and DM to extract and learn relevant features. As a result, 3D CCVFM helps to learn more relevant pose discrimination features, substantially improving network accuracy and achieving higher pose recognition performance.

b: STRONG INTERACTION PROPERTY AMONG VIEWS

The Convolutional Neural Network (CNN) boasts three kinds of principal features: intra-channel, inter-channel, and channel fusion. The convolutional layer of CNN facilitates interaction between channels and generates new channels in the subsequent layer. If there is no view fusion function present, Multi-View Convolutional Neural Networks (MVCNN) struggle to aggregate discriminative features. This challenge arises from the mutual independence of inception multi-view data among views fed into the network. The absence of a view fusion function will impede the detection algorithm from effectively learning and exploiting the complementary information across views. This paper proposes a pivotal operation in the 3D CCVFM that extends dimensionality of the feature map, enabling a focus on the information interaction of the novel dimension (the view dimension), transforming its size from $\mathrm{H} \times \mathrm{W} \times \mathrm{C}$ to $\mathrm{H} \times \mathrm{W} \times \mathrm{N}_{\mathrm{view}}^2 \times \mathrm{C}$. The most remarkable aspect is the $1 \times 1 \times \mathrm{N}_{\mathrm{view}}^2 \times \mathrm{C}$ convolution operation in 3D CCVFM, which conducts view-versus-view and channel-versus-channel interactions, as opposed to interactions within the same view and channel. This unique interaction mode plays an integral role in disentangling the intricate relationship between features among channels and strengthening the connection among diverse views.

2) QUANTITATIVE EVALUATION

a: SPARSENESS EVALUATION

To validate that the feature maps processed by 3D CCVFM aggregate more critical features with a discriminatory effect, we analyzed the output feature maps of multiple FRMs and 3D CCVFM from the perspectives of sparsity and Euclidean distance. Moreover, we employed the sparseness function defined by Hoyer [39] to measure and evaluate the feature map x. This function is widely acknowledged and accepted for assessing the sparseness of neural representations.

$$\mathrm{S}(x) = \frac{\sqrt{n} - \left(\sum |x_i|\right) / \sqrt{\sum x_i^2}}{\sqrt{n} - 1} \tag{11}$$

n in (11) refers to the dimensionality of the feature map x. The function evaluated by $\mathrm{S}(x)$ returns a value of unity if and only if x contains only a single non-zero component. Conversely, it takes a value of zero if all components of x are non-zero and their absolute values are equal. The formula for $\mathrm{S}(x)$ suggests that its value is smaller for smoother images, and larger for images with more texture. It should be noted that the sparseness of x, as measured by $\mathrm{S}(x)$, lies between 0 and 1, with lower values indicating lower sparsity. This relationship is clearly demonstrated in Fig. 7, which presents the statistical results of our experiments on the test set.

To be more explicit, based on the feature response maps (FRMs) of each view, the output feature maps exhibit a sparseness distribution ranging from approximately 0.44 to 0.47, whereas the output of 3D CCVFM demonstrates a sparseness level of around 0.19. Results from the experiments indicate a discernible reduction in feature map sparsity for all test samples. This reduction in sparsity can be attributed to the 3D CCVFM method aggregating a greater number of discriminating features.

b: EUCLIDEAN DISTANCE EVALUATION

Conventionally, the similarity between two images has been measured by computing the Euclidean distance between their corresponding feature representations [40]. This method of similarity calculation is convenient and widely used in various domains such as image retrieval [40], [41], semantic labeling [42], face recognition [43], [44], motion capture [45], clustering algorithms [46], [47], and others, owing to its effectiveness and ease of computation. In the last few years, scholars have introduced more sophisticated techniques for evaluating image similarity, such as deep learning-based models [48]. Nevertheless, despite the advent of these innovative approaches, the Euclidean distance measure persists as a pertinent and commonly employed tool for gauging image similarity, thanks to its straightforwardness and comprehensibility.

In this paper, the Euclidean distance method is employed to address the distance between two vectors, which can range from 0 to infinity. The degree of similarity increases as the distance between the vectors decreases. To this end, (12)

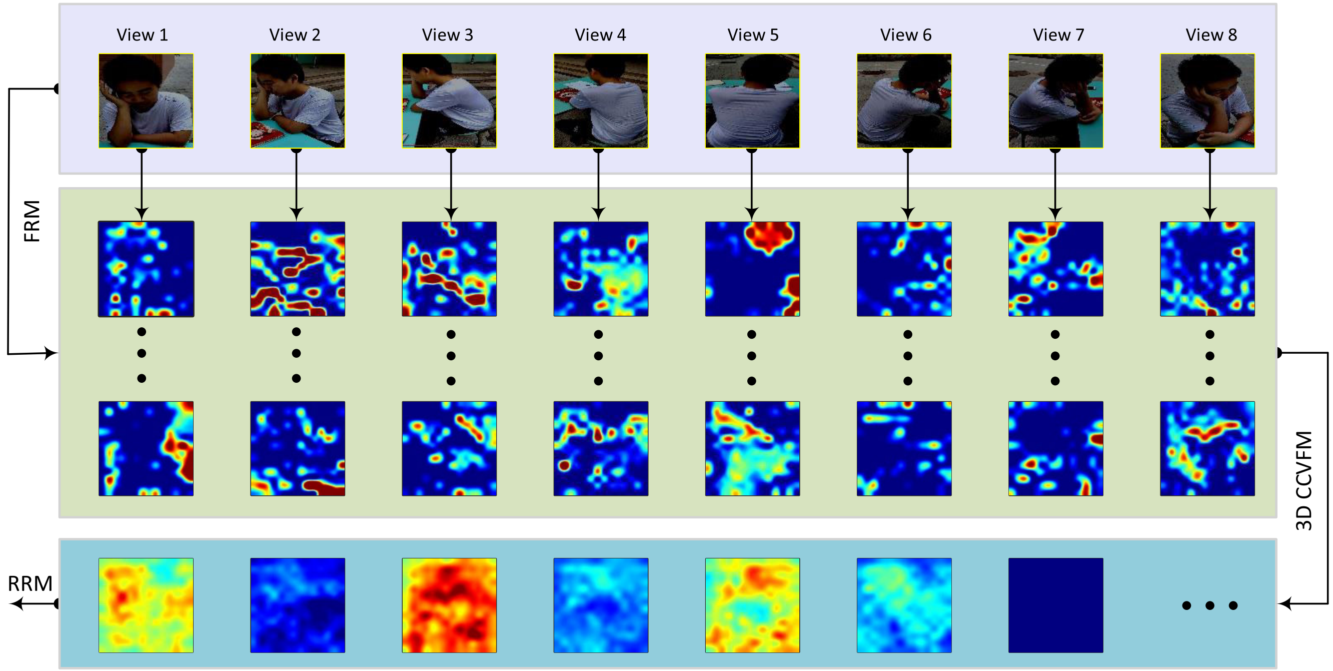

FIGURE 6. Visualization of FRM and 3D CCVFM using Channel Arrangement Method. Each small square heatmap in the figure represents a channel or a node in the input or output of 3D CCVFM. The channel arrangement method is used to spread the input and output along the channels for visualization purposes.

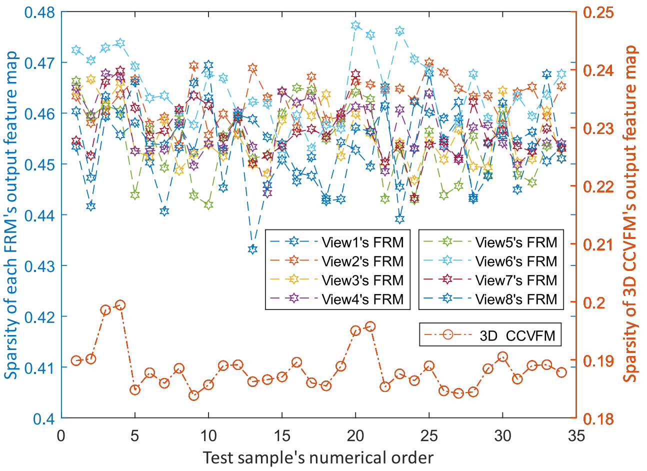

FIGURE 7. Experimental results and statistics of Sparsity. The outputs of FRM are corresponding to the left blue vertical axis and the outputs 3D CCVFM is corresponding to the right brown vertical axis.

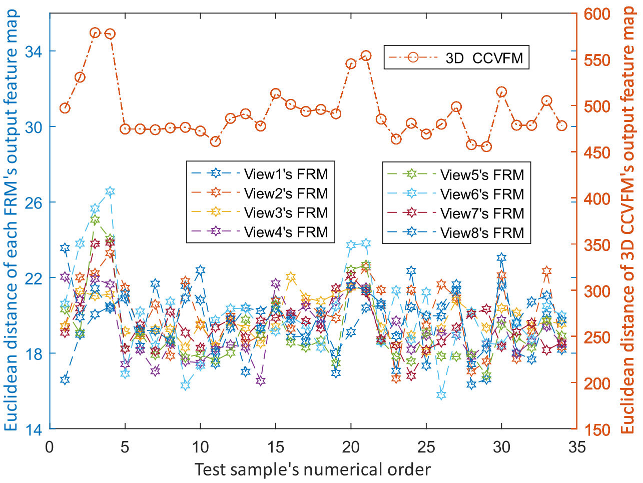

FIGURE 8. Experimental results and statistics of Euclidean distance. The outputs of FRM are corresponding to the left blue vertical axis and the outputs 3D CCVFM is corresponding to the right brown vertical axis.

is utilized to calculate the Euclidean distance between any two channels in the feature map. Subsequently, the Average Euclidean Distance (AED) is obtained by taking the average of the $c^2 - c$ preliminary calculation values (where c represents the number of feature vector channels), which can be used as a measure for evaluating the feature map.

$$D = \sqrt{\sum\nolimits_{i=1}^{n} (x_i - y_i)^2} \tag{12}$$

$$\mathcal{L}_{\text{AED}} = \text{Avg}(D_1, \ldots, D_{c^2-c}) \tag{13}$$

D denotes the preliminary calculation Euclidean distance values, with x and y representing two diverse channels, and n denoting the dimensionality of x and y. The mean evaluation value of the final calculation is represented by $\mathcal{L}_{\text{AED}}$. Fig. 8 demonstrates the experiment statistical result.

The implementation of the 3D CCVFM has significantly enhanced the Average Euclidean Distance (AED) of the feature map, from approximately 20 to around 500. This results in an order of magnitude in the overall level of AED, indicating that 3D CCVFM focuses on strengthening the discriminative ability of the feature map among channels. Subsequent fusion of the features leads to a reduced similarity among channels, an increased discrimination degree, and improved identification accuracy.

In summary, the 3D CCVFM enhances the discriminative features of multiple views and reduces the sparsity of feature maps. This module offers several advantages to the network: Firstly, it effectively suppresses irrelevant features and consolidates discriminative features from multiple views, which helps to focus on significant features. Secondly, the 3D CCVFM reduces the number of zero components in feature maps. Thirdly, 3D CCVFM introduces a novel dimension (the ''view'' dimension) of feature maps that promote interaction among different views and strengthen their connections. Lastly, this module improves the Average Euclidean Distance of feature maps by an order of magnitude, thereby increasing the feature discrimination among channels. Overall, the utilization of 3D CCVFM in the network provides significant benefits, including improved feature discrimination and reduced sparsity in feature maps.

V. CONCLUSION

In this study, a novel network for recognizing pseudo 3D human poses from multiple views is proposed. Through comprehensive ablation studies on a self-built dataset, the results demonstrate that the integration of the 3D CCVFM and Pourer Layer significantly improves various performance indicators of the model. Qualitative and quantitative evaluations are conducted to explore the effective principles of the 3D CCVFM from multiple perspectives, validating the effectiveness of our proposed solution. These findings can inform future research in the area of multiple view fusion, and contribute to the development of more robust and accurate recognition models.

ACKNOWLEDGMENT

The authors would like to thank Yefan Cai for support and assistance concerning the work of this manuscript. They would also like to thank Yuchen Guo and Qiuchen Han for many helpful discussions.

REFERENCES

[1] X. Li, Z. Fan, Y. Liu, Y. Li, and Q. Dai, ''3D pose detection of closely interactive humans using multi-view cameras,'' *Sensors*, vol. 19, no. 12, p. 2831, Jun. 2019, doi: 10.3390/s19122831.

[2] L. Chen, H. Ai, R. Chen, Z. Zhuang, and S. Liu, ''Cross-view tracking for multi-human 3D pose estimation at over 100 FPS,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3276–3285, doi: 10.1109/CVPR42600.2020.00334.

[3] J. Dong, Q. Fang, W. Jiang, Y. Yang, Q. Huang, H. Bao, and X. Zhou, ''Fast and robust multi-person 3D pose estimation and tracking from multiple views,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6981–6992, Oct. 2022, doi: 10.1109/TPAMI.2021.3098052.

[4] Z. Zhang, C. Wang, W. Qin, and W. Zeng, ''Fusing wearable IMUs with multi-view images for human pose estimation: A geometric approach,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2197–2206, doi: 10.1109/CVPR42600.2020.00227.

[5] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis, ''Harvesting multiple views for marker-less 3D human pose annotations,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1253–1262, doi: 10.1109/CVPR.2017.138.

[6] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, ''Cross view fusion for 3D human pose estimation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4341–4350, doi: 10.1109/ICCV.2019.00444.

[7] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[8] M. A. Takalkar, S. Thuseethan, S. Rajasegarar, Z. Chaczko, M. Xu, and J. Yearwood, ''LGAttNet: Automatic micro-expression detection using dual-stream local and global attentions,'' *Knowl.-Based Syst.*, vol. 212, Jan. 2021, Art. no. 106566, doi: 10.1016/j.knosys.2020.106566.

[9] S. Pehlivan and P. Duygulu, ''3D human pose search using oriented cylinders,'' in *Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops, ICCV Workshops*, Sep. 2009, pp. 16–22, doi: 10.1109/ICCVW.2009.5457722.

[10] N. Jammalamadaka, A. Zisserman, and J. C. V., ''Human pose search using deep networks,'' *Image Vis. Comput.*, vol. 59, pp. 31–43, Mar. 2017, doi: 10.1016/j.imavis.2016.12.002.

[11] N. Jammalamadaka, A. Zisserman, M. Eichner, V. Ferrari, and C. V. Jawahar, ''Video retrieval by mimicking poses,'' in *Proc. 2nd ACM Int. Conf. Multimedia Retr.*, Jun. 2012, pp. 1–8, doi: 10.1145/2324796.2324838.

[12] H. Wang, P. He, N. Li, and J. Cao, ''Pose recognition of 3D human shapes via multi-view CNN with ordered view feature fusion,'' *Electronics*, vol. 9, no. 9, p. 1368, Aug. 2020, doi: 10.3390/electronics9091368.

[13] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, ''Multi-view convolutional neural networks for 3D shape recognition,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953, doi: 10.1109/Iccv.2015.114.

[14] M. J. Er, Y. Zhang, N. Wang, and M. Pratama, ''Attention pooling-based convolutional neural network for sentence modelling,'' *Inf. Sci.*, vol. 373, pp. 388–403, Dec. 2016, doi: 10.1016/j.ins.2016.08.084.

[15] C. Wang, M. Pelillo, and K. Siddiqi, ''Dominant set clustering and pooling for multi-view 3D object recognition,'' in *Proc. Brit. Mach. Vis. Conf.*, 2017, pp. 1–12, doi: 10.5244/C.31.64.

[16] L. Li, S. Zhu, H. Fu, P. Tan, and C. Tai, ''End-to-end learning local multi-view descriptors for 3D point clouds,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1916–1925, doi: 10.1109/Cvpr42600.2020.00199.

[17] Y. Feng, Z. Zhang, X. Zhao, R. Ji, and Y. Gao, ''GVCNN: Group-view convolutional neural networks for 3D shape recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 264–272, doi: 10.1109/CVPR.2018.00035.

[18] T. Yu, J. Meng, and J. Yuan, ''Multi-view harmonized bilinear network for 3D object recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 186–194, doi: 10.1109/CVPR.2018.00027.

[19] X. Wei, R. Yu, and J. Sun, ''View-GCN: View-based graph convolutional network for 3D shape analysis,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1847–1856, doi: 10.1109/CVPR42600.2020.00192.

[20] D. Shi and H. Tang, ''Research on safe driving evaluation method based on machine vision and long short-term memory network,'' *J. Electr. Comput. Eng.*, vol. 2021, pp. 1–13, Apr. 2021, doi: 10.1155/2021/9955079.

[21] Y. Zhang, J. Li, Y. Guo, C. Xu, J. Bao, and Y. Song, ''Vehicle driving behavior recognition based on multi-view convolutional neural network with joint data augmentation,'' *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4223–4234, May 2019, doi: 10.1109/Tvt.2019.2903110.

[22] A. E. Orhan and X. Pitkow, ''Skip connections eliminate singularities,'' in *Proc. ICLR*, Vancouver, BC, Canada, May 2018, pp. 1–22.

[23] A. Veit, M. Wilber, and S. Belongie, ''Residual networks behave like ensembles of relatively shallow networks,'' in *Proc. NIPS*, vol. 29, Oct. 2016, pp. 1–9.

[24] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, ''Identity mappings in deep residual networks,'' in *Proc. ECCV*, vol. 9908. Amsterdam, The Netherlands, Oct. 2016, pp. 630–645, doi: 10.1007/978-3-319-46493-0_38.

[25] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, ''Aggregated residual transformations for deep neural networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5987–5995, doi: 10.1109/CVPR.2017.634.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[27] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, ''Deep networks with stochastic depth,'' in *Proc. ECCV*, vol. 9908. Amsterdam, The Netherlands, Oct. 2016, pp. 646–661, doi: 10.1007/978-3-319-46493-0_39.

[28] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, ''Dual path networks,'' in *Proc. NIPS*, Long Beach, CA, USA, Dec. 2017, pp. 1–11.

[29] S. Targ, D. Almeida, and K. Lyman, ''ResNet in ResNet: Generalizing residual architectures,'' in *Proc. ICLR*, SAN Juan, Puerto Rico, May 2016, pp. 1–7.

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[31] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, ''Dropout: A simple way to prevent neural networks from overfitting,'' *J. Mach. Learn. Res.*, vol. 15, pp. 1929–1958, Sep. 2014.

[32] X. Li, W. Wang, X. Hu, and J. Yang, ''Selective kernel networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519, doi: 10.1109/CVPR.2019.00060.

[33] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, ''ResNeSt: Split-attention networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1–22.

[34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, ''Squeeze-and-excitation networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

[35] M. Zhao, S. Zhong, X. Fu, B. Tang, and M. Pecht, ''Deep residual shrinkage networks for fault diagnosis,'' *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4681–4690, Jul. 2020, doi: 10.1109/Tii.2019.2943898.

[36] M. Lin, Q. Chen, and S. Yan, ''Network in network,'' in *Proc. ICLR*, Mar. 2014, pp. 1–10.

[37] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, ''Gradient-based learning applied to document recognition,'' *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998, doi: 10.1109/5.726791.

[38] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, ''Learning spatiotemporal features with 3D convolutional networks,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497, doi: 10.1109/ICCV.2015.510.

[39] P. O. Hoyer, ''Nonnegative matrix factorization with sparseness constraints,'' *J. Mach. Learn. Res.*, vol. 5, no. 9, pp. 1457–1469, Nov. 2004.

[40] X. Chen and Y. Li, ''Deep feature learning with manifold embedding for robust image retrieval,'' *Algorithms*, vol. 13, no. 12, p. 318, Dec. 2020, doi: 10.3390/a13120318.

[41] P. Liu, L. Shi, Z. Miao, B. Jin, and Q. Zhou, ''Relative distribution entropy loss function in CNN image retrieval,'' *Entropy*, vol. 22, no. 3, p. 321, Mar. 2020, doi: 10.3390/e22030321.

[42] P. Nguyen, K. Nguyen, R. Ichise, and H. Takeda, ''EmbNum: Semantic labeling for numerical values with deep metric learning,'' in *Proc. JIST*, Awaji, Japan, Nov. 2018, pp. 119–135, doi: 10.1007/978-3-030-04284-4_9.

[43] R. Gao, F. Yang, W. Yang, and Q. Liao, ''Margin loss: Making faces more separable,'' *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 308–312, Feb. 2018, doi: 10.1109/Lsp.2017.2789251.

[44] R. Goel, I. Mehmood, and H. Ugail, ''A study of deep learning-based face recognition models for sibling identification,'' *Sensors*, vol. 21, no. 15, p. 5068, Jul. 2021, doi: 10.3390/s21155068.

[45] J. Sedmidubsky, P. Elias, and P. Zezula, ''Effective and efficient similarity searching in motion capture data,'' *Multimedia Tools Appl.*, vol. 77, no. 10, pp. 12073–12094, May 2018, doi: 10.1007/s11042-017-4859-7.

[46] L. Jin, X. Wang, J. Chu, and M. He, ''Human activity recognition machine with an anchor-based loss function,'' *IEEE Sensors J.*, vol. 22, no. 1, pp. 741–756, Jan. 2022, doi: 10.1109/Jsen.2021.3130761.

[47] Q. Ji, Y. Sun, J. Gao, Y. Hu, and B. Yin, ''Nonlinear subspace clustering via adaptive graph regularized autoencoder,'' *IEEE Access*, vol. 7, pp. 74122–74133, 2019, doi: 10.1109/Access.2019.2920592.

[48] S. Chen, C. Guo, and J. Lai, ''Deep ranking for person re-identification via joint representation learning,'' *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016, doi: 10.1109/TIP.2016.2545929.

YUANFENG XIE received the B.S. degree from the Wuhan University of Technology, China, in 2020. He is currently pursuing the M.S. degree with the Department of Physics, State Key Laboratory of Optoelectronic Materials and Technologies, Sun Yat-sen University. His current research interests include computer vision and image processing.

XIANGYANG YU received the D.Sc. degree from Sun Yat-sen University China, in 1998. He is currently a Professor and the Ph.D. Director with the Department of Physics, Sun Yat-sen University. He has published more than 100 articles in his research-related fields and obtained multiple invention patents and software copyrights. His major research interests include intelligence processing of optical imaging, deep learning, computer vision, and intelligent photoelectric device.

WEIBIN HONG received the B.S. and M.S. degrees from Sun Yat-sen University, China, in 2015 and 2018, respectively. He is currently involved in algorithm development and model research related to artificial intelligence. His current research interests include machine vision, deep learning, and hyperspectral data processing techniques.

ZHAOLONG XIN received the M.S. degree from the University of Waterloo, Canada, in 2013. He has applied for six invention patents, ten utility model patents, and two software copyrights. His current research interests include intelligent industrial inspection systems, computer vision, and machine learning.

YANWEN CHEN graduated from the Nanjing University of Posts and Telecommunications, China. He is currently a Senior Researcher with Guangzhou Gaoke Communications Technology Company Ltd. His current research interests include machine learning, deep learning, and computer vision.