Multi-view 3D human pose estimation based on multi-scale feature by orthogonal projection

Yinghan Wang, *Jianmin* Dong[*], *Yanan* Wang, and *Bingyang* Sun

College of Information Engineering, Xizang Minzu University,712082 Xianyang Shaanxi, China

Abstract. Aiming at the problems of inaccurate estimation results, complicated matching of feature information in different views and poor robustness of the network model in complex scenes, a multi-view multi-person 3D human pose estimation model with multi-scale feature orthogonal projection is proposed, which includes a multi-scale orthogonal projection fusion network and an orthogonal feature ascending dimension network. Firstly, the multi-scale orthogonal projection fusion network performs orthogonal projection of features at multiple scales, using the residual structure to fuse features in the same plane separately, simplifying the feature learning difficulty and reducing the feature loss due to projection. Then, it is fed into the orthogonal feature ascending dimension network to reconstruct higher level 3D features using trilinear interpolation and deconvolution to improve the expressiveness of the model, and finally fed to the backbone network to supplement the information of the high-dimensional features, and the network regresses according to the different stages of the task to obtain the 3D human pose. The experimental results show that the Percentage of 3D Correct Parts is improved on the Campus and Shelf datasets, and the Mean Per Joint Position Error is reduced on the CMU Panoptic dataset and the average accuracy is improved at a smaller threshold compared to the previous method. The prediction results are also better than the previous method by reducing the perspective input on the trained model. The proposed method not only effectively estimates the 3D human pose, but also improves the prediction accuracy and enhances the robustness of the network model.

1 Introduction

With the continuous development of deep learning, the field of human pose estimation has progressed by leaps and bounds, where 3D human pose estimation is a more challenging and more relevant problem. 3D human pose estimation has a wide range of application scenarios in different fields such as medical rehabilitation[1], human performing arts[2], and human-computer interaction[3]. It can be divided into single-person and multi-person scenes in multi-view pose estimation.

For multi-view single-person 3D pose estimation methods [4~7], the 2D human pose of the view is estimated first [8,9], and then the 3D human pose is estimated by combining the camera parameters of the corresponding perspective [10]. Qiu et al. [4] combined geometric

[*] Corresponding author: jmdong@xzmu.edu.cn

priors from different views to obtain a more accurate 2D human pose, and on the basis of the image structure model the authors proposed a recursive graph structure model, which in turn recovered the 2D human pose as a 3D human pose. In view of the fact that the cost of obtaining accurate 3D human pose data is expensive and some network models require a large amount of labeled 3D human pose data, Kocabas et al. [5] applied self-supervised learning methods to multi-view single-person 3D pose estimation, without camera external parameters, and obtained 3D human pose through multi-view 2D human pose and epipolar geometric constraints. For the problem of poor robustness to different groups of cameras in feature fusion models for multiple views, Xie et al. [6] proposed a pre-training model for meta-learning, which learns from a large scale of cameras by meta-learning, eliminating the need for each group of cameras to train the fusion model separately and improving the adaptability to different cameras. Gong et al. [7] designed an augmentation framework with differentiable operations to learn to adjust the pose, which can be jointly optimized with a 3D pose estimation network to continuously adjust the data augmentation. In addition, a partial perceptual motion chain space is introduced, and a discriminator is designed to ensure the rationality of the generated data, so as to generate more realistic and diverse posture data and improve the generalization of the model. The 3D pose estimation of single and multi-person in multiple views is closely related, but the 3D pose estimation of multiple people in multiple views [11-16] is more complex, involving more information matching and fusion due to the many moving and more occluded characters in the scene, and some excellent methods have recently emerged. Dong et al. [11] proposed an unsupervised method that combines geometric constraints and appearance information to perform 2D pose matching across views, and then uses the 3DPS model to infer the 3D pose of each person based on the matching results, but the method relies on the 2D pose estimation results, and the 2D pose estimation results affect the matching results of the human body in multi-view, which in turn affects the 3D reconstruction results of the human body, and the method is prone to chain effects. To avoid making incorrect matching decisions in each view, Tu et al. [12] proposed a voxel representation-based method to fuse feature information from multiple views by processing 2D heatmaps of the human body from multiple views according to camera parameters and projecting them into a discretized stereoscopic space, but the network model uses a large number of 3D convolutions making the network training more difficult and the estimation results in some scenes are highly biased. Wu et al. [13] designed a Multi-view Matching Graph module to match the centres of the same person in different views, and then used the matching results for simple triangulation to obtain candidate regions for 3D human centroids; the Center Refinement Graph module was used to search for candidate points in a spherical range to further refine the centroid positions; the predicted 3D human centres were used with a fixed size 3D bounding box and a 3D pose estimator to generate the initial 3D pose, and the authors proposed a human Pose Regression Graph model to further refine the initial 3D pose, but the process was more tedious. Lin et al. [14] proposed a plane sweep stereo-based approach with cross-view consistency constraints by a plane sweep algorithm to perform accurate depth regression and finally reconstruct the 3D human pose based on the estimated depth information, but the authors reduced the number of predicted joints and reduced the difficulty of model prediction. Chu et al. [15] initialised the 3D pose by crossing the 2D correspondence of the views. The authors calculated the affinity of the 2D pose to the 3D pose by considering only joints with positive affinity, used temporal consistency to match the 2D pose of different views with the previously constructed 3D pose, and used limit constraints to remove noise, and finally obtained a more accurate 3D pose. Ye et al. [16] followed the voxel representation by performing orthogonal projection of features in discrete space, performing 2D-CNNs or 1D-CNNs learning only on the three mutually orthogonal 2D feature maps obtained, estimating partial joint coordinates from each of the three planes before fusing them thus obtaining the final 3D pose, which is lower than the previous network model in terms of prediction accuracy for both common datasets, although it improves the run rate.

Although the above methods achieve good results, some of them [11,13, 15] require tedious matching tasks during implementation, which can easily lead to making wrong matching decisions and thus affect the estimation of 3D human pose; some methods [14] estimate a relatively small number of key points of the human body and do not show the 3D human pose well. voxelpose [12] and Faster-voxelpose [16] models are good at avoiding the tedious matching task, but the prediction results in complex scenes are poor and the network model is less robust in the case of reduced number of cameras. The Voxelpose model uses a large number of 3D convolutions, which increases the training difficulty and memory consumption. To address these problems, this paper combines the respective advantages of the Voxelpose [12] and Faster-voxelpose [16] models and also adopts the form of voxel representation, by constructing a multi-scale orthogonal projection fusion network, realising the downscaling of 3D features to 2D features, obtaining a more compact and effective feature representation, and combining high-level semantic information with low-level detail information learning, allowing the network to learn based on 2D orthogonal features, solving the problem of sparse and difficult to learn feature points in discrete 3D space; by constructing an orthogonal feature ascending dimension network, more and richer 3D feature information is recovered based on the fused deep 2D orthogonal features and using their mutually corroborating properties. By combining 2D and 3D features in this way for feature diversity learning, the accuracy and robustness of the 3D pose estimation task is improved.

2 Methods

In complex scenes, the interaction of the human body is random and the feature points are relatively sparse in 3D space. Sometimes the human body will have different degrees of occlusion, which makes human pose estimation more difficult. Our network is based on a two-stage learning strategy, from coarse to fine, and alternately performs orthogonal projection and feature learning on the three scale features in the backbone network. The projected planes are surface xoy, surface xoz, and surface yoz, respectively, and the three planes are orthogonal to each other.

The task of the first stage of the mesh is to roughly locate the human in space. We set the size of the stereo space to 8m×8m×2m, and set the adjacent grid spacing to 100mm and discretize it, because in most cases this space size is sufficient to cover most human activities, and the 100mm spacing is also sufficient to roughly locate the human in space in the first stage. The task of the first stage of the mesh is to roughly locate the human in space. We set the size of the stereoscopic space to 8m×8m×2m, and the adjacent grid interval to 100mm, and discretize it, because this space size is sufficient to cover most human activities in most cases, and the 100mm interval distance is also sufficient to roughly locate the human in space in the first stage. The features of the discretized three-dimensional space are fed into our network, and the backbone network to roughly locate the 3D position of the human body in space. We feed three different scales of features from the backbone network into the multi-scale orthogonal projection fusion network, perform coarse two-dimensional learning on the three mutually orthogonal planar features, and then use the orthogonal feature ascending network to ascend and fuse the human location information from the three orthogonal features and feed it into the backbone network. For the initial feature projected into stereoscopic space, a small network is slipped, and the slider window centered on the anchor point is mapped to a low-dimensional feature, which is then sent to the fully connected layer to regress the confidence score of the location and output the confidence score of the location for all anchor points that have a human body likelihood. We calculate a ground truth heatmap score based on the distance between each anchor and the ground truth location of the root joint, and the score decreases as the distance increases. The locations where the score is greater than a threshold are selected by NMS, which in turn is used to extract local peaks. This step optimizes the network by calculating the root mean square loss of the network output with respect to the ground truth heatmap score.

The second stage shares a network with the first stage. In contrast to the first stage, the network reproduces a more detailed 3D human pose based on the rough localization of the first stage. At this stage, we set the size of the stereoscopic space to 2000mm×2000mm×2000mm and divide it into discrete bodies of 64×64×64, because setting such a space size is enough to cover various poses of a human body. Unlike the first stage, the projection here is personal, mapping a grid in 3D space onto a 2D image plane, and then obtaining a heatmap projection of a single human body based on the rough positioning of the human body. We feed the stereo spatial features of a single human body into the backbone network to regress the detailed 3D pose information of the human body. The effect of 2D human pose representation varies in different orthogonal planes. In order to fully learn the feature semantic information, we use a multi-scale orthogonal projection fusion network to regress the 2D human pose information in different planes, and use the orthogonal feature ascending network to aggregate the 2D pose information in three planes to recover the 3D pose information and feed it into the backbone network. The delineated discrete mesh still has a large error in the pose regression of a single person, we have reduced the error by regressing the 3D heatmap of the key points of the human body and calculating its center of mass to obtain the 3D position of each key point of the human body by calculating the expectation. This step optimizes the network based on the L1 loss between the calculated 3D positions of the human key points and the Ground-truth positions. The structure of this network is shown in Figure 1.

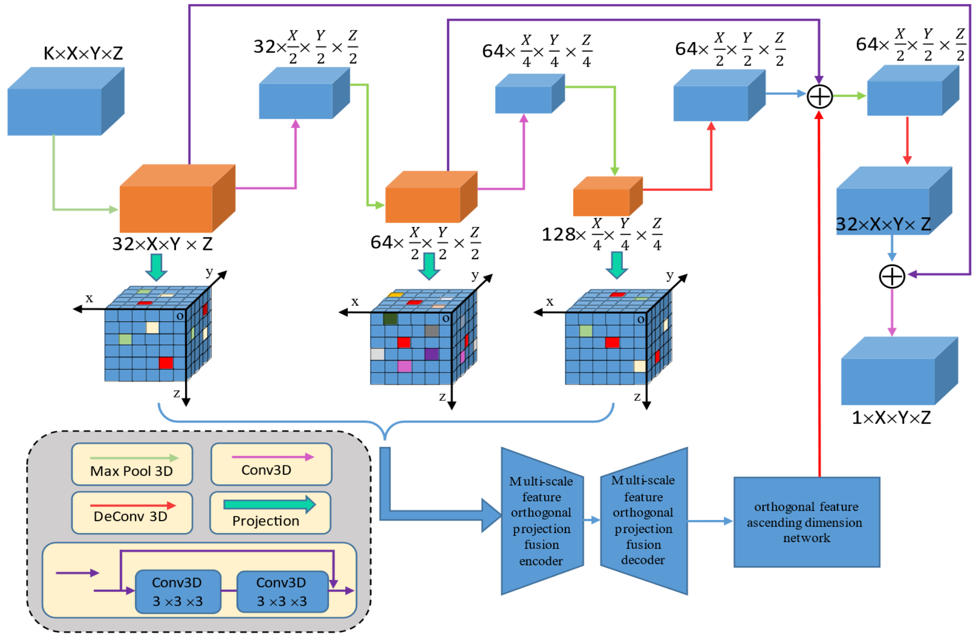


Fig. 1. Network model structure.

2.1 Multi-scale orthogonal projection fusion network

The backbone network of our network model removes some network layers and heavy jump connections compared to Voxelpose [12], because stacking too many network layers may make the effective depth of the network insufficient; during the network training process, a large number of 3D convolution and 3D residual structures make the network occupy more video memory resources and make training more difficult, etc. In this regard, we feed three different scales of 3D features extracted during network model learning into the multi-scale orthogonal projection fusion network to learn the semantic information of 2D human pose, thus reducing the difficulty of feature learning and enriching the feature information with this multi-dimensional feature learning.

We perform convolution and pooling operations on the initial features $V_0 \in \Re^{X \times Y \times Z}$ projected into the discrete space, and extract three features I_1, I_2, I_3 at different scales for orthogonal projection, which are computed as follows:

$$I_1 = \sigma(f^{7\times7\times7}(V_0)) \tag{1}$$

$$I_2 = \sigma(f^{3\times3\times3}(f_{\max}(I_1))) \tag{2}$$

$$I_3 = \sigma(f^{3\times3\times3}(f_{\max}(I_2))) \tag{3}$$

where $f^{7\times7\times7}(\cdot)$, $f^{3\times3\times3}(\cdot)$ are convolutions with convolution kernel size 7×7×7, 3×3×3, respectively. $f_{\max}(\cdot)$ is the 3D global Max pooling operation. $\sigma(\cdot)$ is the batch normalization, ReLU activation function operation on the convolution data. Among them, the step size in each convolution operation is 1, and no padding operation is performed.

We project the features I_1, I_2, and I_3 of the backbone network from shallow to deep into three mutually orthogonal planes: plane xoy, plane xoz, and plane yoz, respectively. The orthogonal projection operation is performed by maximizing the corresponding coordinate axes, and the multidimensional tensor is projected along a certain coordinate axis by taking the maximum value of the compressed projection to obtain the projection in the three mutually orthogonal planes. The procedure is as follows:

$$P_m^{xoy}(I_m)_{i,j,k,1} = max_{n=1}^{d}(I_m^{i,j,k,n}) \tag{4}$$

$$P_m^{xoz}(I_m)_{i,j,1,l} = max_{n=1}^{w}(I_m^{i,j,n,l}) \tag{5}$$

$$P_m^{yoz}(I_m)_{i,1,k,l} = max_{n=1}^{h}(I_m^{i,n,k,l}) \tag{6}$$

where P_m^{xoy} is the result of the projection of the tensor I_m onto the plane xoy, c is the number of channels, and h, w, d is the dimension of the tensor in the x,y,z directions, $i \in [1,c]$, $j \in [1,h]$, $k \in [1,w]$, $l \in [1,d]$, $m \in [1,3]$, $I_m \in \Re^{c\times h\times w\times d}$.

Although voxel-based representation [12,16] avoids making wrong decisions in each view, voxels are generally very sparse in stereo space, leading to many unnecessary computations. Although we achieve dimensionality reduction by orthogonal projection to reduce the redundant information and thus the difficulty of feature learning, it involves a loss of data information. By projecting the high-dimensional features onto three orthogonal planes, we can recover a lot of information through their mutually corroborating properties, and we can compensate for the information loss to some extent by re-fusing the feature projections at multiple scales. We sequentially orthogonally project the low-level features onto the high-level features of the backbone network, and fuse the projected features of different scales in the same plane to learn the geometric relationships of the human body planes. The multi-scale orthogonal projection fusion network is shown in Figure 2.

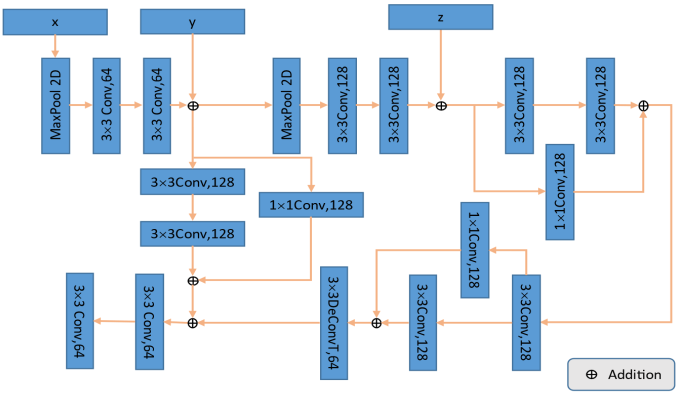


Fig. 2. Multi-scale orthogonal projection fusion network.

This network is constructed using an encoding-decoding approach that uses residual blocks instead of ordinary convolution to extract high-quality feature information. Inspired by the Resent network and its powerful feature learning capability, we construct a multi-scale orthogonal projection fusion network using residual blocks. We perform orthogonal projection of features at three different scales, each scale is projected onto three mutually orthogonal planes, and the features at different scales in the same plane are fused early after projection to allow the network to learn information of different features more comprehensively and reduce the loss of information between features at different scales. To obtain richer semantic information of features, we try a step-by-step feature fusion strategy to incorporate same-scale features at different stages of the network. First, the input features were globally max-pooled, and then the features with the same projection scale as I_2 were obtained through a residual block. Again, features I_1 and I_2, projected on the same plane, were added to increase their information. Among them, the same is true for the projection fusion process from I_2 to I_3. The feature fusion coding process is described as follows.

$$\varphi_{12}^{q} = F(MaxPool(P_1^q)) + Conv_{3\times3}(MaxPool(P_1^q)) + P_2^q \tag{7}$$

$$\varphi_{23}^{q} = F(MaxPool(\varphi_{12}^q)) + Conv_{3\times3}(MaxPool(\varphi_{12}^q)) + P_3^q \tag{8}$$

where P_m^q denotes the result of the orthogonal projection of feature I_m in plane q; φ_{12}^q denotes the result of the initial fusion of features I_1 and I_2 after projecting them to plane q respectively; φ_{23}^q denotes the result of the pre-fusion of the features of I_2 projected onto the plane q with φ_{12}^q, $q \in \{xoy, xoz, yoz\}$; $F(\cdot)$ denotes two convolutional layers without jump connection and with a convolutional kernel size of 3 × 3, with a batch normalization, ReLU activation function operation performed by default after each convolutional layer; a residual block is denoted as $\varphi = F(x) + x$.

The features after orthogonal projection were sent to the multi-scale orthogonal projection fusion network, the features of different scales in the same plane were fused, and the final three orthogonal plane features were aggregated, and the obtained 3D features were sent to the backbone network decoder of the same scale for feature fusion by dimension elevation. Our network combines 3D features and 2D features well. The 3D network is used to learn the spatial geometry of the human body, and the 2D network is used to learn the planar geometry of the human body. The backbone network decoder mainly recovers more details of voxel space features by deconvolution, and then completes the feature reconstruction operation in stereo space.

2.2 Orthogonal feature ascending dimension network

The network up-dimensions and aggregates the fused orthogonal projection features in turn, and the fused three features are still orthogonal, and we up-dimension the three orthogonal features to the same size by up-sampling trilinear interpolation. Based on the mutual corroboration of the three orthogonal features, relatively reliable 3D feature information can be extracted, so the three orthogonal features are important to us, and in view of this, we up-dimensionalize the three orthogonal features and use the averaging strategy to express the 3D features after fusing the multi-scale orthogonal projections. In order to extract the high-level semantic information of the upscaled 3D features, we give the upscaled features to a convolutional layer for learning, and let this layer convolutionalize to make the feature extraction strategy, using deconvolution [17] a convolutional kernel of size 3×3×3 with a step size of 1 to aggregate the feature information again, and this phase of the network is shown in Figure 3.

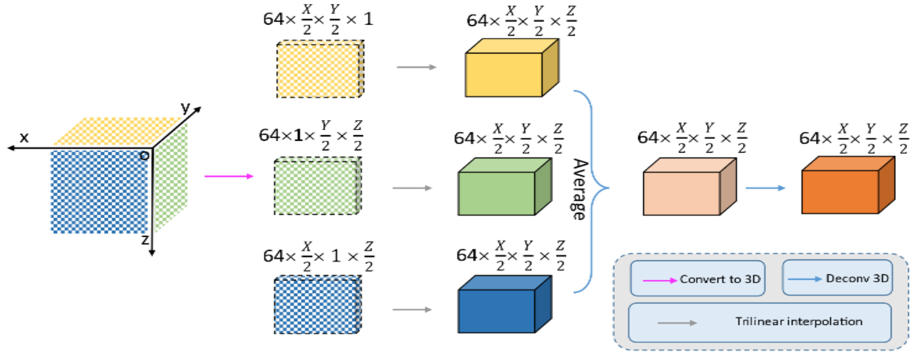

$64\times\frac{X}{2}\times\frac{Y}{2}\times 1$
$64\times\frac{X}{2}\times\frac{Y}{2}\times\frac{Z}{2}$
y
x
o
z
$64\times 1\times\frac{Y}{2}\times\frac{Z}{2}$
$64\times\frac{X}{2}\times\frac{Y}{2}\times\frac{Z}{2}$
Average
$64\times\frac{X}{2}\times\frac{Y}{2}\times\frac{Z}{2}$
$64\times\frac{X}{2}\times\frac{Y}{2}\times\frac{Z}{2}$
$64\times\frac{X}{2}\times 1\times\frac{Z}{2}$
$64\times\frac{X}{2}\times\frac{Y}{2}\times\frac{Z}{2}$
Convert to 3D
Deconv 3D
Trilinear interpolation

Fig. 3. Orthogonal feature ascending dimension network.

For I_1, I_2, I_3, three features are projected on three orthogonal planes xoy, xoz, yoz, and the fused results are E^{xoy}, E^{xoz}, E^{yoz}. At this time, the fused results are 2D, we construct the pseudo-3D shape, and then use the upsampling trilinear interpolation method to increase the dimension of the features to the specified space size. The process is expressed as follows:

$$\begin{gathered}H(x,y,z)=(1-u)(1-v)(1-w)H(x_0,y_0,z_0)+u(1-v)(1-w)H(x_0,y_0,z_0)\\+v(1-u)(1-w)H(x_0,y_1,z_0)+vu(1-w)H(x_1,y_1,z_0)\\+w(1-u)(1-v)H(x_0,y_0,z_1)+vw(1-u)H(x_0,y_1,z_1)+uvwH(x_1,y_1,z_1)\end{gathered}\tag{9}$$

$$u=\frac{x-x_0}{x_1-x_0},v=\frac{y-y_0}{y_1-y_0},w=\frac{z-z_0}{z_1-z_0}\tag{10}$$

where u , v , and w are the interpolation coefficients between interpolation point (x, y, z) and the last eight data points, and $H(x_i, y_j, z_k)$ is the value of the data point at point (x_i, y_j, z_k) .

$$\psi^{xoy} = H(E'_{xoy}),\ \ \psi^{xoz} = H(E'_{xoz}),\ \ \psi^{yoz} = H(E'_{yoz}) \tag{11}$$

where E'_{xoy} , E'_{xoz} , E'_{yoz} represent E^{xoy} , E^{xoz} , E^{yoz} pseudo 3D shape; $H(\cdot)$ denotes the up-sampled trilinear interpolation function. ψ^{xoy} , ψ^{xoz} , ψ^{yoz} denote the features according to the dimension elevation.

3 Experiments

3.1 Setting of experimental environment and related parameters

The experimental environment is under Windows10 system, Pytorch1.12.1, CUDA11.6, Python3.7, and NVIDIA GeForce RTX 3090 GPU. On the Campus and Shelf datasets, the size of the input images in the training stage is 800×640 and 800×608, the number of human keypoints is 17, the heat map score threshold is 0.1, the number of iterations is 50, and the detection space size is 12000mm×12000mm×2000mm. On the CMU Panoptic dataset, the input image size is 960×512, the number of human keypoints is 15, the heatmap score threshold is 0.3, the number of iterations is 20, and the detection space size is 8000 mm×8000 mm×2000mm. On these three datasets, the BatchSize is set to 4, the initial cube size is set to 80×80×20, and the regression single person pose cube size is 64×64×64.

3.2 Comparison experiment and visualization of experimental results

The results of our comparison with other models on the Campus and Shelf datasets [18] under the PCP3D metric are shown in Tables 1 and 2. From the results in Table 1, it is shown that on the Campus dataset, our method improves 0.1%, 0.6%, and 6.2%, respectively, compared to the methods of Tu et al [12], Ye et al [2], and Ershadi-Nasab et al [20] for the PCP3D metric, and compared to them, our method improves 1.4%, 2.5%, and 4. 8% on Actor1; The results through Table 2 show that on Shelf dataset the method of this paper improves 1.2%, 0.6%, and 10.2% in PCP3D metrics compared to the methods of Tu et al [12], Ye et al [2], and Ershadi-Nasab et al [22], and on Actor1, Actor2 the method of this paper outperforms the results of the previous methods. The experiments on Campus and Shelf datasets show that our method achieves good results by discarding a large amount of 3D convolution and compensating for the loss of feature semantic information by using multi-scale orthogonal projection fusion networks and orthogonal feature boosting networks. Although the annotation of the data in the Campus and Shelf datasets is limited, the method in this paper can still predict the 3D human pose well and achieves better results compared to previous methods.

Table 1. Comparison of PCP3D results of different methods on the Campus dataset.

Method	Actor 1	Actor 2	Actor 3	Average
Belagiannis et al.[3]	82.0	72.4	73.7	75.8
Belagiannis et al.[18]	93.5	75.7	84.4	84.5

Ershadi-Nasab et al.[20]	94.2	92.9	84.6	90.6
Ye et al.[16]	96.5	94.1	97.9	96.2
Wang et al.[21]	98.2	94.1	97.4	96.6
Tu et al.[12]	97.6	**93.8**	**98.8**	96.7
Ours	**99.0**	93.7	97.9	**96.8**

Table 2. Comparison of PCP3D results of different methods on the Shelf dataset.

Method	Actor 1	Actor 2	Actor 3	Average
Belagiannis et al.[3]	66.1	65.0	83.2	71.4
Belagiannis et al.[18]	75.3	69.7	87.6	77.5
Ershadi-Nasab et al.[20]	93.3	75.9	94.8	88.0
Ye et al.[16]	99.4	96.0	97.5	97.6
Wang et al.[21]	99.3	95.1	**97.8**	97.4
Tu et al.[12]	99.3	94.1	97.6	97.0
Ours	**99.5**	**97.3**	97.7	**98.2**

In Fig. 4, we visualize the results of the proposed method on the Campus dataset, where the red color is the corresponding true data annotation, and the other colors are the predicted results of the proposed method. In the first case, the coarse localization of the human center almost matches the truth, and the coarse localization is already sufficient for our regression task; for the 3D human pose with truth values, our prediction results are highly consistent with the truth values. Among the three views, the man in the black shirt only appears in the first view, and is completely invisible in the second and third views without truth labels. In the second case, the man in black still only appears in one view, and in the second view, one actor almost completely obscures the other actor. In this challenging scene, our model still predicts the approximate location of the human body center and the full 3D human pose well.

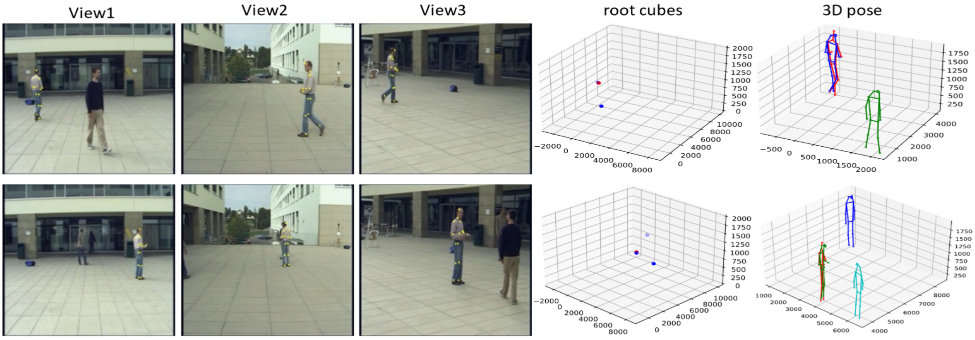


Fig. 4. Visualization results on the Campus dataset.

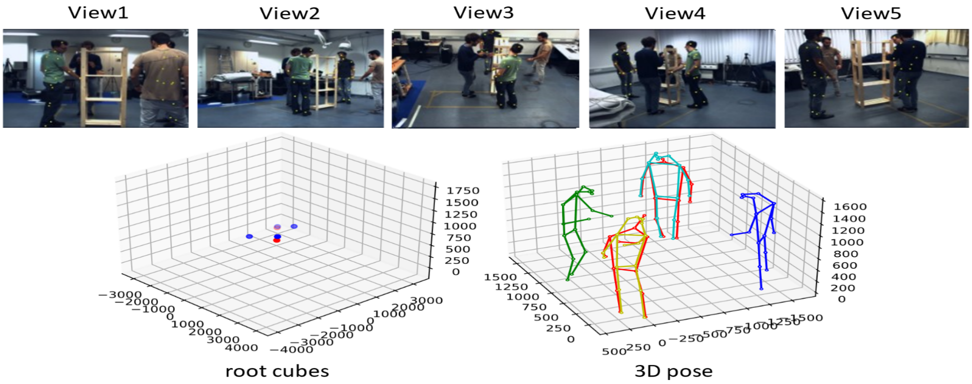


Fig. 5. Visualization results on the Shelf dataset.

In Fig. 5, the results of the proposed method on the Shelf dataset are visualized, where red is the corresponding true data annotation, and other colors are the predicted results of the proposed method. The human pose data annotation in the Shelf dataset is also incomplete, the number of views increased to four, and the number of active people in the scene increased to four, not only there is occlusion between people, but the shelf in the scene also occludes part of the human body, which brings some difficulty to the prediction of 3D human pose. From the comparison of the visual prediction results and the true value results, the proposed method predicts the human pose very well.

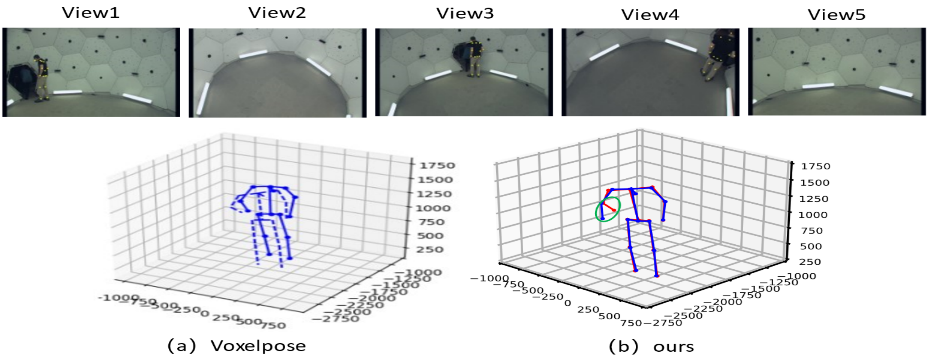
On the CMU Panoptic dataset[19], our model is compared with other models in AP and MPJPE evaluation metrics, where AP is the average precision and is the PR curve area. In the PR curve, the horizontal axis is the proportion of correctly identified human joint positions, that is, the recall rate. The vertical axis is the proportion of positive samples that are truly positive, which is the precision. The accuracy and robustness of the model were evaluated by calculating the area under the PR curve; higher values of PR indicate better model performance, and the comparison results are shown in Table 3.

Table 3. Comparison results of different models on CMU Panoptic dataset.

Method	AP_{25}	AP_{50}	AP_{100}	AP_{150}	MPJPE
Voxelpose[12]	83.59	98.33	**99.76**	**99.91**	17.68mm
Faster-voxelpose[16]	85.22	98.08	99.32	99.48	18.26mm
Ours	**86.20**	**98.38**	99.66	99.77	**17.57mm**

The experimental results on the CMU Panoptic dataset show that the MPJPE index of the proposed method is 0.10 mm and 0.69 mm lower than those of Voxelpose [12] and Fast-VoxelPose [16], respectively. Our AP_{25} is increased by 2.61% and 0.98%, respectively, indicating that the method in this paper has a better performance of the network model under a more stringent threshold.

In the Voxelpose [12] network model, the authors give a scenario where their method predicts poorly on the CMU Panoptic dataset, and to better demonstrate the advantages of the method in this paper, we visualize the results in the same scenario with the Voxelpose [12] network model. The blue solid line in Fig. 6 shows the model prediction results, the blue dashed line in Fig. 6(a) shows the true value results, and the red solid line in (b) shows the true value results. From the visualized results, we can see that the predicted 3D human pose of the Voxelpose [12] network model in this scene shows a large deviation, while our predicted results are highly consistent with the true value results, with only a little deviation.



(a) Voxelpose (b) ours

Fig. 6. Comparison results on the CMU Panoptic dataset.

3.3 Robustness verification

We further compare with Voxelpose [12] and Faster-voxelpose [16] models, and train the corresponding models on the CMU Panoptic dataset [18] using data from five viewpoints. The trained model evaluates the MPJPE index by decreasing the number of cameras in turn under the same test set. It can be seen from the results that our method has a slight increase compared to Voxelpose[12] when the number of cameras is 4. Under other numbers of cameras, the MPJPE index of our method is significantly smaller than that of other methods. Although the model is trained under five views, by reducing the number of cameras to predict the 3D pose of the human body, the prediction results of our method gradually open a gap with other methods, indicating that the robustness of our method is better than other methods. The comparison results are shown in Fig. 7.

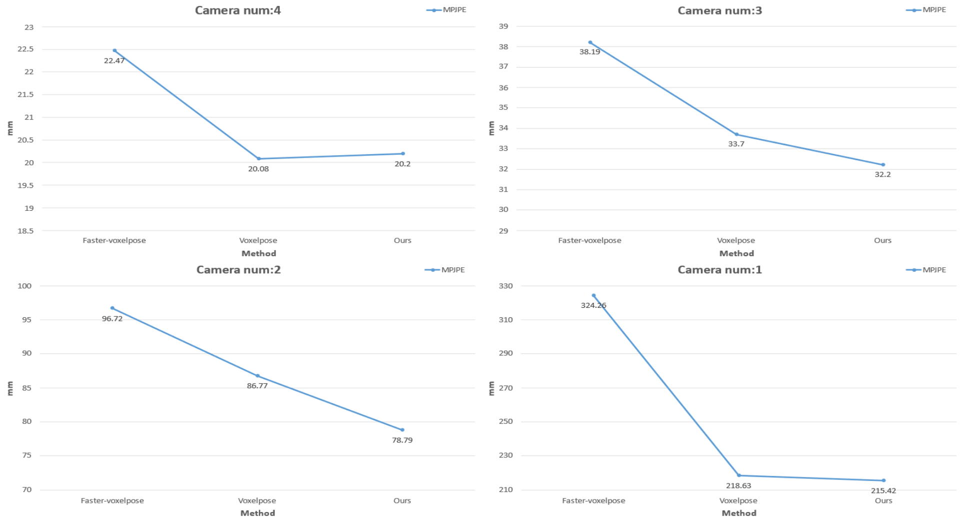


Fig. 7. The results of different trained models under a certain number of cameras.

This work is supported by the 2022 Scientific Research Project of Xizang Minzu University(22MDY013).

References

1. H. -C. Shih ,A survey of content-aware video analysis for sports, IEEE Transactions on Circuits and Systems for Video Technology, **28.5** (2017): 1212-1231.
2. R. Li, S. Yang, D.A. Ross, et al. Learn to dance with aist++: Music conditioned 3d dance generation. arXiv preprint arXiv:2101.08779, 2021, 2(3).
3. Song, Yale, David Demirdjian, and Randall Davis, Continuous body and hand gesture recognition for natural human-computer interaction, ACM Transactions on Interactive Intelligent Systems, **2.1** (2012): 1-28.
4. H. Qiu, C. Wang, J. Wang, et al. Cross view fusion for 3d human pose estimation, *Proceedings of the IEEE/CVF International Conference on Computer Vision*. (2019): 4342-4351.
5. M. Kocabas, S. Karagoz, E. Akbas, Self-supervised learning of 3d human pose using multi-view geometry, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019): 1077-1086.
6. R. Xie, C. Wang, Y. Wang, Metafuse: A pre-trained fusion model for human pose estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020): 13686-13695.
7. K. Gong, J. Zhang, J. Feng, Poseaug: A differentiable pose augmentation framework for 3d human pose estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021): 8575-8584.
8. S. Wang, K. Hang, ZH. Chen, et al. Survey on 3D Human Pose Estimation of Deep Learning, Journal of Frontiers of Computer Science and Technology, **17.1** (2023):74-87.
9. K. Sun, B. Xiao, D. Liu, et al. Deep high-resolution representation learning for human pose estimation, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2019): 5693-5703.
10. Y. Peng, J. Guo, CH. Yu, et al. Calibration method for high precision camera based on plane transformation, Journal of Beijing University of Aeronautics and Astronautics, **48.7** (2022):1297-1303.
11. J. Dong, W. Jiang, Q. Huang, et al. Fast and robust multi-person 3d pose estimation from multiple views, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* (2019): 7792-7801.
12. H. Tu, C. Wang, W. Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment, *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer International Publishing*, (2020): 197-212.
13. S. Wu, S. Jin, W. Liu, et al. Graph-based 3d multi-person pose estimation using multi-view images, *Proceedings of the· IEEE/CVF International Conference on Computer Vision*, (2021): 11148-11157.
14. J. Lin, G. H. Lee, Multi-view multi-person 3d pose estimation with plane sweep stereo, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021): 11886-11895.
15. H. Chu, J. Lee, Y. Lee, et al. Part-aware measurement for robust multi-view multi-human 3d pose estimation and tracking, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2021): 1472-1481.

16. H. Ye, W. Zhu, C. Wang, et al. Faster VoxelPose: Real-time 3D Human Pose Estimation by Orthographic Projection, *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI. Cham: Springer Nature Switzerland,* (2022): 142-159.
17. A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. Springer International Publishing*, (2016): 483-499.
18. V. Belagiannis, S. Amin, M. Andriluka, et al. 3d pictorial structures revisited: Multiple human pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, **38.10** (2015): 1929-1942.
19. H. Joo, H. Liu, L. Tan, et al. Panoptic studio: A massively multiview system for social motion capture, *Proceedings of the IEEE International Conference on Computer Vision*, (2015): 3334-3342.
20. S. Ershadi-Nasab, E. Noury, S. Kasaei, et al. Multiple human 3d pose estimation from multiview images. Multimedia Tools and Applications, **77** (2018): 15573-15601.
21. T. Wang, J. Zhang, Y. Cai, S. Yan, et al. Direct multi-view multi-person 3d pose estimation[J]. Advances in Neural Information Processing Systems, **34** (2021): 13153-13164.