Journal of Beijing Institute of Technology, 2026, Vol. 35, No. 2

MRM: Multi-View 3D Human Pose Estimation Based on Regression with Multivariate Joint Distribution

Anzhan Liu✉, Hufei Zhao✉, Yilu Ding
(School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China)

Abstract: In the field of multi-view three-dimensional (3D) human pose estimation, there are primarily two approaches: heatmap-based and regression-based models. Regression-based models require less computational effort than heatmap-based models but are less accurate. This study proposes a regression-based model called multi-view 3D human pose estimation based on regression with multivariate joint distribution (MRM), which achieves accuracy comparable to heatmap-based models while using lower computational resources in multi-view 3D human pose estimation. Specifically, this model employs a flow-based method to learn the multivariate joint distribution of human pose data, enabling the regression-based model to capture nonlinear dependencies across different perspectives. Experimental results on two public datasets validate the accuracy and efficiency of the proposed model. Compared with heatmap-based methods, MRM reduces multiply-add operations by 32.3% while maintaining comparable prediction accuracy.

Keywords: human pose estimation; log-likelihood estimation; convolutional network; multivariate joint distribution

1 Introduction

Human pose estimation refers to detecting the positions of various body parts of the human body in images or videos. The task of predicting the coordinate positions of two dimensions is referred to as two-dimensional (2D) human pose estimation, whereas predicting the coordinate positions in three dimensions is termed three-dimensional (3D) human pose estimation. The latter requires predicting the depth information of the human pose. Human pose estimation is a crucial issue in computer vision. Nowadays, human pose estimation can be integrated into various applications, including virtual reality, human-computer interaction, and human medical rehabilitation detection.

As research on 2D human pose has achieved excellent results, the study of 3D human pose estimation has attracted much attention. Compared to 2D human pose estimation, 3D human pose estimation suffers from inherent ambiguity in image depth information. Therefore, it is even more challenging. Due to occlusion issues in human images, such as self-occlusion and environmental occlusion, relying solely on images from a single perspective can result in the loss of important information. This results in a reduction in the accuracy of the model. Many existing studies focus on monocular 3D human pose estimation, which struggles to effectively handle occlusions. Therefore, this work is dedicated to multi-view human pose estimation research. It can complement severely occluded parts through images from multiple angles.

Manuscript received Sep. 11, 2025; revised Nov. 5, 2025; accepted Dec. 15, 2025. The associate editor coordinating the review of this manuscript was Dr. Lijuan Jia.
✉ Corresponding author. Email: liuanzhan@zut.edu.cn, hufeizhao@163.com
DOI: 10.15918/j.jbit1004-0579.2025.061

Multi-view human pose estimation faces two major challenges that urgently need to be addressed. On the one hand, most mainstream methods are based on heatmaps, which demand a large amount of computational and storage resource. On the other hand, while regression-based methods are concise, they struggle with integrating information from multiple views and exhibit low prediction accuracy. Applying multi-view human pose estimation to real-life scenarios inevitably faces the issue of real-time performance. In terms of converting image representation information into numerical coordinates of the target, it can be divided into two categories: one is the heatmap-based method, and the other is the regression-based method. The heatmap-based method is the mainstream approach in this field. This method involves inputting images into a convolutional neural network. By applying convolutional and deconvolutional operations, probability heatmaps are generated for each keypoint of the human body. In these heatmaps, each pixel represents the probability that the corresponding location in the image is a human keypoint. The positions of the human keypoints are then identified through the argmax operation. Due to its additional deconvolution process and the fact that its result accuracy is proportional to the heatmap resolution, the heatmap-based method requires a significant amount of computational and memory resource. When this method is applied to three-dimensional multi-view scenarios, the resource requirements escalate further, making it challenging to meet real-time demands. The regression-based method directly maps the input image to joint coordinates for output. It is efficient and concise, with low demands on computational resources. With the same backbone network architecture, the regression head of multi-view 3D human pose estimation based on regression with multivariate joint distribution (MRM) consumes only 7.6% of the multiply-add operations compared to the heatmap head. Therefore, it can easily meet the needs of real-time applications. However, by fusing heatmaps, models based on heatmaps [1] can obtain two-dimensional coordinates simultaneously utilizing information from other views. In contrast, during multi-view fusion, the information provided by different perspectives is often complementary and unique (for example, one perspective captures the frontal details of an object, while another reveals the occluded side). Regression models tend to learn and output an "average" or "the most probable" value, which can blur or even discard the unique information offered by multiple perspectives. As a result, regression models struggle to effectively address the challenges of multi-view fusion.

Furthermore, previous works only adopted traditional L_1 (least absolute deviations / Manhattan norm) or L_2 (least squares / Euclidean norm) losses. [2, 3] showed that using L_1 or L_2 losses actually assumes that the human pose distribution satisfies the Laplace distribution or Gaussian distribution. However, the true distribution of human pose data does not conform to either of these. Therefore, the prediction accuracy of the model decreases.

To address the aforementioned problems and challenges, the methodology of conducting the literature review in this study followed the systematic procedure put forward by Saied et al. [4] We proposed a regression model, MRM, which can learn the multivariate joint distribution of multi-view human poses. Data from different viewpoints often exhibit strong complementary information. These multi-view data can be regarded as manifestations of the same pattern from different orientations or levels. MRM learns the relationships among different perspectives through a multivariate joint distribution. In this way, the model can comprehensively utilize multi-view human pose information. Unlike previous works that only use traditional L_1 or L_2 losses, we used a normalizing flow model [5–7] to learn the

multivariate joint distribution of multi-view human pose data through maximum likelihood estimation. During model training and inference, this distribution serves as significant prior knowledge to enhance the model's prediction accuracy. Considering the complexity of learning the multivariate joint distribution of human pose data, we used a normalizing flow model to learn the deviation between the target distribution and common initial distributions, such as a multivariate normal distribution. This enhancement significantly improves the method's feasibility. Since poses from different viewpoints are not independent, we replaced the zero elements in the covariance matrix of the initial distribution to strengthen the interdependencies among various dimensions of the multivariate distribution, making the training process easier to optimize. To better integrate multi-view information, MRM learns the weights of each joint from multi-view images. The learned weights can balance the importance of each joint in recovering the 3D coordinates.

The regression-based approach for obtaining coordinates has low computational resource requirements. During the training phase, the regression process for the multivariate joint distribution of human poses and the three-dimensional coordinates of human joints can be performed synchronously, without requiring additional training steps.

In summary, our contributions are as follows:

We proposed a multi-view regression model to predict 3D pose information, significantly reducing the computational and storage requirements of the model, in order to satisfy the real-time needs of edge devices.

We enhanced the model's prediction accuracy by utilizing a flow model to learn the multivariate joint distribution of multi-view human poses through maximum likelihood estimation and employing joint weights to co-train images from various perspectives.

We conducted experiments on human 3.6 million (Human3.6M) [8] (a large-scale dataset for 3D human pose estimation in various actions) and Max Planck Institute for Informatics - 3D human pose in the wild (MPI_INF_3DHP) [9] (another dataset for 3D human pose estimation, focusing on more challenging and diverse scenarios) to verify the accuracy and efficiency of the model.

2 Related Works

2.1 Multi-View 3D Pose Estimation

Multi-view 3D pose estimation typically involves fusing multi-view features, such as voxel aggregation [10–12] and geometric constraints [13, 14], and then predicting the 3D coordinates of human joints using deep learning methods. During training, relying solely on 2D keypoint information to predict 3D information is often underconstrained. Chun et al. [15] addressed this issue by developing an autoencoder-based voxel heatmap representation learning method to learn human prior knowledge. This method innovatively uses voxel heatmaps as constraints and exhibits strong advantages in cross-dataset generalization. Iskakov et al. [10] aggregated synchronized images from multiple views into a 3D voxel using projection matrices based on camera parameters and obtained 3D heatmaps through a voxel network, achieving excellent results on the Human 3.6M dataset. However, the model requires calculating multi-view 2D heatmap aggregation and subsequent 3D heatmaps, resulting in considerable computational cost. To address this issue, Remelli et al. [16] proposed a lightweight multi-view human pose estimation method that obtains a camera-view-independent feature by applying a linear transformation module to the learned 2D heatmaps, eliminating the need for voxel aggregation and improving the model's computational speed. He et al. [13] introduced the Epipolar Transformer, which fuses features from different views in the intermediate layers of a 2D detec-

tion network, enabling the utilization of 3D information during 2D detection and thereby improving its accuracy. Ma et al. [14] observed that joint predictions from one view are related to the image not only on the epipolar line in another view, but also to its surrounding areas. Therefore, they proposed the concept of epipolar field to learn more comprehensive correlated features, which is an extension of the epipolar line [13]. They also combined it with the positional encoding of Transformer [17] to encode the 3D relationships among different views. Zhang et al. [18] presented a simple and effective multi-view feature fusion method. This method utilizes the geometry of perimeters to determine point-to-line correspondences among different views. Then, it implicitly determines point-to-point correspondences through the sparse nature of heatmaps, avoiding complex point-to-point matching processes. In this paper, we modeled the relationship among human poses from different viewpoints to comprehensively utilize multi-view information.

2.2 Human Pose Estimation Based on Regression

In human pose estimation, heatmap-based methods remain the primary approach. Although heatmap-based methods are able to preserve the spatial information of images and achieve high prediction accuracy, they face issues such as the inability to train end-to-end and quantization problems that affect accuracy. Moreover, heatmap-based methods have a relatively high computational cost. Regression-based methods, on the other hand, are more concise as they do not require generating heatmaps. Carreira et al. [19] proposed a novel framework that introduces structured prior knowledge through feedback to better address structured prediction problems. Liang et al. [20] proposed using bones instead of joint points to represent human poses, unifying 2D and 3D pose estimation, and enabling end-to-end training using both 2D and 3D training data. Li et al. [2] proposed using flow models to learn the underlying distribution of human pose outputs to enhance the prediction accuracy of regression models. Unlike their approach, we aimed to learn the correlations among human pose data from various views through normalizing flow models.

Martinez et al. [21] first utilized an efficient 2D pose estimation model to obtain 2D human joint points and then employed regression to lift the 2D coordinates to 3D coordinates. Although sometimes different 3D points may project to the same 2D point, introducing ambiguity in this lifting process, this method still has advantages as it can leverage the off-the-shelf 2D human pose estimation models for efficient 2D joint prediction. In single-view prediction, the state-of-the-art model [22] also adopted this approach and proposed to disrupt real 3D poses by gradually adding Gaussian noise, and then trained a denoising network to restore them, generating multiple 3D pose hypotheses. It further proposed a joint-level reprojection-based multi-hypothesis aggregation method, joint-level reprojection-based multi-hypothesis aggregation (JPMA), to aggregate the generated hypotheses in a reasonable manner. Zhu et al. [23] proposed using a unified framework to address various human-centered video tasks, employing a pre-training-fine-tuning framework to learn a general human motion representation. They proposed utilizing the lifting of 2D skeleton sequences to 3D skeletons as a self-supervised pre-training task. However, these two-stage methods still inevitably utilize 2D heatmaps when obtaining 2D pose information through existing 2D human pose estimation models. Our proposed model does not utilize 2D heatmaps and directly predicts joint coordinates through regression.

3 Method

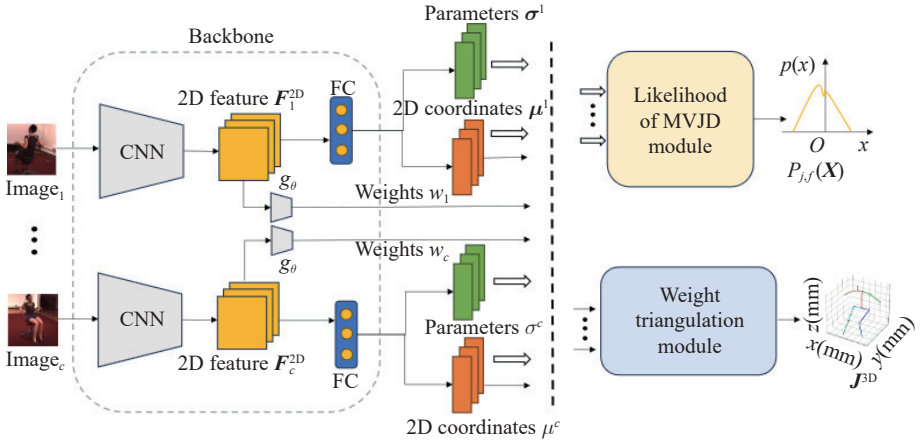
3.1 Overview of the Method

In this work, we proposed a regression model with log-likelihood for multi-view human pose

estimation. The goal is to evaluate the 3D joint coordinates of humans $\{j_i^{3D} = (x_i, y_i, z_i) \,|i \in \{1, 2, \cdots, J\}\}$ in the world coordinate system from multiple synchronized images, where J denotes the number of human joints. These synchronized multi-view images are captured from C calibrated cameras. The complete process is illustrated in Fig. 1, which comprises a backbone network, a likelihood of multivariate joint distribution (MVJD) module, and a weight triangulation module. The backbone network first extracts image features $\boldsymbol{F}_c^{2D} \in \mathbf{R}^{K \times H \times W}$ and joint weights $W_c \in \mathbf{R}^J$ from the input multi-view images $\boldsymbol{I}_c \in \mathbf{R}^{H_0 \times W_0 \times 3}$. The multi-view image features $\boldsymbol{F}_c^{2D}$ are passed through fully connected layers to output normalized 2D coordinates of human joints $\{\mu_i^c = (x_i^c, y_i^c) \,|i \in \{1, 2, \cdots, J\}\}, \boldsymbol{\mu}^c \in \mathbf{R}^{J \times 2}$ and distribution parameters $\boldsymbol{\sigma}^c \in \mathbf{R}^{J \times 2}$. The combination of $\boldsymbol{\mu}^c$ and $\boldsymbol{\sigma}^c$ is then input into the likelihood of MVJD module to learn the multivariate joint distribution of multi-view human poses. Meanwhile, $\boldsymbol{\mu}^c$ and W_c are fed into the weight triangulation module, where linear algebra triangulation is utilized to obtain the 3D joint coordinates $\boldsymbol{J}^{3D} \in \mathbf{R}^{J \times 3}$. The details will be discussed in the following sections.

3.2 Backbone

We combined synchronized images from different views as a single input. Using the ground truth bounding boxes, we cropped each input image frame and fed it into convolutional networks to obtain intermediate 2D features $\{\boldsymbol{F}_c^{2D}\}_{c=1}^C$ and joint weights $\{W_c\}_{c=1}^C$. Our convolutional network is mainly composed of residual network (ResNet)-152 [24], which is pre-trained on the common objects in context (COCO) [25], MPII human pose (MPII) [26] and Human3.6M [8]. Unlike the general pose ResNet, MRM does not require generating heatmaps, thus eliminating the process of deconvolution for features $\boldsymbol{F}_c^{2D}$. This significantly reduces the number of model parameters as well as the training and detection time. The joint weights W_c are obtained by inputting the features $\boldsymbol{F}_c^{2D}$ into a convolutional network g_θ. It consists of two convolutional layers, a max pooling layer, and three fully connected layers, where θ denotes the learnable parameters of the convolutional network. Then, the features $\boldsymbol{F}_c^{2D}$ are passed through a subsequent average pooling layer to reduce the size of the feature map to 1. After that, they are fed into a fully connected layer to obtain $\boldsymbol{\mu}^c$ and $\boldsymbol{\sigma}^c$. To enhance the model's generalization ability, normalization processing was conducted on $\boldsymbol{\mu}^c$ and $\boldsymbol{\sigma}^c$, where $\boldsymbol{\mu}^c$ is normalized by dividing the L_2 norm of the input features, and $\boldsymbol{\sigma}^c$ is normal-



Note: The red-green-blue (RGB) images are taken as the input for the model. The likelihood of MVJD module learns the multivariate joint distribution of multi-view human poses and the weight triangulation module recovers the 3D human pose. CNN represents convolutional neural network. FC represents fully connected layer.

Fig. 1 Overview of our MRM model

ized using a sigmoid function.

3.3 Likelihood of MVJD Module

Likelihood of MVJD refers to the likelihood of multivariate joint distribution. Li et al. [2] leveraged normalizing flow models to learn the underlying distribution of human poses. However, it models observations from each viewpoint independently. Inspired by [2], we did not model the images from multiple perspectives individually. While this preserves the unique information provided by each perspective, it weakens the interconnections among different views. To address this, we proposed using a multivariate joint distribution to characterize the relationships among multi-view data and employed a normalizing flow model (such as real-valued non-volume preserving (Real NVP) [5]) to learn this multivariate joint distribution. This normalizing flow model can achieve a reversible and stable mapping between data space and latent space by scale and translation operation. And it is trained by maximizing the log-likelihood. We suggested consulting the original paper for additional information for more details.

The process is shown in Fig. 2. It is difficult to visualize high-dimensional joint distributions in a two-dimensional plane. We used a univariate distribution as an approximation to represent the multivariate joint distribution. First, we initialized a distribution, which can be an arbitrary simple distribution. Here, we used an initial normal distribution $Z \sim N(0, I)$ whose covariance matrix is an identity matrix. Directly applying a flow model f_ϕ (where ϕ represents the learnable parameters of the flow model) to map and transform this distribution into $\overline{\boldsymbol{X}} \sim P_\phi(\overline{\boldsymbol{X}})$ clearly does not meet our expectations. This is because there must be some underlying relationships among the multi-view perspectives of the pose, while the identity matrix implies that the variables are mutually independent—since all off-diagonal elements of the identity matrix are zero. To address this, we set all off-diagonal elements to 0.5. The resulting initialized distribution is then transformed using the flow model, which simplifies the transformation process.

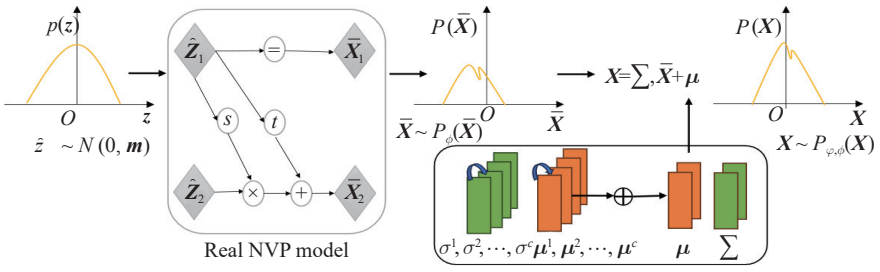
We transformed the data dimensions, as shown in Fig. 2, by swapping the joint dimension and the viewpoint dimension. This is because we need to capture the relationships between the same poses across different perspectives. The joint dimension is input into the flow model as the first dimension. And concatenate various viewpoint dimensions to obtain $\boldsymbol{\mu} \in \mathbf{R}^{J\times 2c}$, $\boldsymbol{\sigma} \in \mathbf{R}^{J\times 2c}$. In other words, each input consists of the pose information of a specific human joint from various camera viewpoints.

This operation was performed to capture the relationships of the same pose across different perspectives.

Using the obtained $\boldsymbol{\sigma}$ and $\boldsymbol{\mu}$, we applied translation and scaling to the distribution variables to derive the final distribution

$$\boldsymbol{X} = \boldsymbol{\sigma}\overline{\boldsymbol{X}} + \boldsymbol{\mu} \tag{1}$$

According to the density formula of random variable transformation, our final distribution is



Note: s and t stand for scale and translation operation. φ represents the learnable parameters in the backbone. ♁ stands for the concatenate operation.

Fig. 2 The architecture of likelihood of MVJD module

$$P_{\varphi,\phi}(\boldsymbol{X}) = P_{\phi}(\overline{\boldsymbol{X}}) \left| \det \left(\frac{\partial \overline{\boldsymbol{X}}}{\partial \boldsymbol{X}} \right) \right| \quad (2)$$

The loss function is

$$L_{\text{mle}} = -\ln P_{\varphi,\phi}(\boldsymbol{X}) = -\ln P_{\phi}(\overline{\boldsymbol{X}}) + \ln \boldsymbol{\sigma} \quad (3)$$

In the calculation of model inference, it's important to note that MRM requires only parameters $\boldsymbol{\mu}$ to infer the 3D coordinates from the weight triangulation module. The likelihood of MVJD module does not need to participate in inference. As a result, this exclusion of the likelihood of MVJD module further reduces the computational cost of the model.

3.4 Weight Triangulation Module

We used weight triangulation module (WTM) to convert the 2D coordinates of pose joints into the predicted $\boldsymbol{J}^{3D}$ coordinates. Specifically, the predicted 2D coordinates are transformed into 3D coordinates using the method of non-homogeneous linear algebraic triangulation. This requires solving a characteristic equation system in three-dimensional coordinates: $\boldsymbol{A}_i \boldsymbol{j}_i = 0$ to obtain the predicted 3D coordinates $\boldsymbol{J}^{3D}$. Here, $\boldsymbol{A}_i \in \mathbf{R}^{2C\times 4}$, $\boldsymbol{A}_i$ is a matrix constructed from the projection matrices and the 2D joint coordinates. The system of equations is derived by solving the inverse projection relationship $\boldsymbol{\mu}' = \boldsymbol{M}\mathrm{J}^{3D}$, where $\boldsymbol{M} \in \mathbf{R}^{3\times 4}$ denotes the camera's projection matrix. The prediction accuracy for the same joint point typically varies across different perspectives. If the joint coordinates $\boldsymbol{\mu}'$ with different accuracies are used with the same weight to infer the 3D joint coordinates $\boldsymbol{J}^{3D}$, it can lead to model degradation, especially when predicting complex poses. Therefore, we introduced W to dynamically balance the noisy 2D observation information from different perspectives, resulting in

$$(\boldsymbol{W}_i \times \boldsymbol{A}_i)\, \boldsymbol{j}_i = 0 \quad (4)$$

where $\times$ denotes element-wise multiplication. Performing singular value decomposition on the equation system yields

$$(\boldsymbol{W}_i \times \boldsymbol{A}_i)\, \boldsymbol{j}_i = \boldsymbol{U} \square \boldsymbol{D} \square \boldsymbol{V}^{\mathrm{T}} \quad (5)$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices, and $\boldsymbol{D}$ is a diagonal matrix containing the singular values. The loss function is

$$L_{\text{mse}} = \frac{1}{N} \sum\nolimits_{n=1}^{N} (\boldsymbol{J}^{3D} - J^{\text{gt}})^2 \quad (6)$$

where N and J represent the total number of joint points and the ground truth coordinates, respectively. J^{gt} represents the ground truth joint coordinates of the human body. Combining the two losses yields the final loss function

$$L_{\text{total}} = L_{\text{mle}} + \lambda L_{\text{mse}} \quad (7)$$

where λ is loss weight.

4 Experiments

4.1 Datasets

Human3.6M [8] is a large-scale dataset for 3D human pose estimation, which includes 3.6×10^6 frame images and 3D human joint labels. It was captured by four synchronized, calibrated cameras, with each camera operating at a frequency of 50 Hz. The dataset involves 11 professional actors performing in 17 action scenarios. Following protocol 1 [21, 27], we used (S 1, S 5, S 6, S 7, S 8) for training and (S 9, S 11) for validation. During training and validation, each action is processed independently.

The MPI_INF_3DHP [9] dataset is suitable for 3D human pose estimation. It was acquired through 14 cameras in an indoor green-screen studio. The dataset contains 8 subjects, each performing 8 activities. The 8 activities are divided into two sequences, namely sequence 1 and sequence 2. Sequence 1 includes walking/standing, exercise, sitting 1, and crouch/reach. Sequence 2 comprises on the floor, sports, sitting 2, and miscellaneous. The official test set of this dataset only includes single-view images, while its training set contains multi-view images. Therefore, we split its training set into a training subset and a test subset. Consistent with pre-

vious work [28], we chose S 1 – S 7 for training and S 8 for testing. To better compare with Human3.6M, we adopted images captured by 4 cameras as input, namely cameras 0, 2, 7, and 8. Since the dataset does not include bounding box information for the subjects, we used the maximum and minimum values of the ground truth 2D joint coordinates along the x and y axes as the bounding box coordinates.

4.2 Evaluation Metrics

We used the mean-per-joint-position-error (MPJPE) as an evaluation metric for 3D pose estimation, which calculates the average of the Euclidean distances between the predicted 3D joint coordinates and the ground truth values. For each human pose, there are typically multiple joint points (such as the head, shoulders, elbows, wrists, hips, knees, ankles, etc.), and each joint point corresponds to a 3D coordinate (x, y, z). By summing up the errors of all joint points and being divided by the total number of joint points, we obtained the mean joint error. Our prediction involves 17 joints of the human body. The MRM model is able to directly predict the joint coordinates in the world coordinate system, so there is no need to perform subsequent rigid body transformations to align the predicted joints with the ground truth when calculating the Euclidean distances.

MPJPE focuses on the absolute error of individual joint positions, whereas 3D percentage of correct keypoints (3DPCK) places greater emphasis on evaluating the overall accuracy of a model's predictions for human joint points. 3DPCK is another metric used to assess the accuracy of 3D human pose estimation algorithms in predicting joint points. It calculates the distance between predicted joint points and true joint points, and determines whether this distance is less than a preset threshold. If the distance is less than the threshold, the joint point is considered correctly predicted. In this manner, the percentage of correctly predicted joint points out of all joint points is calculated. On the MPI_INF_3DHP dataset, following the approach [9, 28], we computed the area under the curve (AUC) for a range of 3DPCK values.

4.3 Implementation Details

During training, four views were adopted, and the input image size (H_0, W_0) was set to (384, 384). The 2D features $\boldsymbol{F}_c^{2D}$ extracted from the backbone convolutional network were primarily obtained from ResNet-152 [24], with a feature map (H, W) size of (12, 12) and 2048 channels. In the real NVP model, the structures of the functions s and t are largely similar, both composed of three linear (fully-connected) layers and two activation layers. Each fully-connected layer is followed by a leaky-rectified linear unit (ReLU) layer. The intermediate dimension of the fully-connected layers is set to 128. The input dimensions of the functions s and t vary according to the number of viewpoints: for 2 viewpoints, the dimension is 4; for 3 viewpoints, it is 6; and for 4 viewpoints, it is 8. The input data dimensions are split using a checkerboard mask, with the checkerboard mask defined as [0, 0, 0, 0, 1, 1, 1, 1] and [1, 1, 1, 1, 0, 0, 0, 0]. The Adam optimizer was used for training, with a batch size of 2. The model mainly consists of a backbone network, a likelihood of MVJD module, and a weight triangulation module, with learning rates of 10^{-4}, 10^{-3}, and 10^{-4}, respectively. The training settings for the backbone network follow previous work [10]. The convolutional network for obtaining joint weights consists of two convolutional layers, one max-pooling layer, and three fully connected layers. The training was conducted for 25 epochs. The training of the model was completed on two real-time raytracing (RTX) 2070 Super graphics cards, taking a total of

Tab. 1 Results of the module ablation experiments

Backbone	Multivariate joint distribution	Joint weight	MPJPE (mm)
ResNet-152	×	×	29.21
ResNet-152	×	√	26.39
ResNet-152	√	×	28.20
ResNet-152	√	√	**23.71**

about 2 days.

4.4 Ablation Experiments

The ablation experiment results are shown in Tab. 1. This experiment was conducted on the Human3.6M dataset. We removed the multivariate joint distribution component and used only the mean squared error as the loss function. After removal, the performance of the MRM model significantly declined. This result demonstrates the crucial role of the likelihood of MVJD module in enhancing the model's performance. By learning the multivariate joint distribution of pose data, this module is capable of capturing the complex correlations and dependencies among different viewpoints, thereby providing the model with more comprehensive and accurate information.

In this section, we removed the joint weighting component. The model directly feeds the coordinates obtained from the backbone network into the WTM module and uses the same weight for each view when predicting 3D coordinates.

We visualized the comparison results, as shown in Fig. 3. The first row exhibits the prediction results of the basic model, where we used the camera projection matrix to project the predicted 3D coordinates onto a 2D plane. The second row displays the prediction results of the MRM model for the same set of images, and the third row presents the ground truth corresponding to the images. Fig. 4 reveals the joint weights learned by the MRM model for each view. As can be clearly seen in the first column referred to as the view (a), the right wrist joint is occluded by the human body, so the prediction of the right wrist by the basic model deviates significantly from the actual value. However, our MRM model assigns a very small weight to joint 10, which represents the right wrist. The model can learn information about the right wrist from other views, thus achieving higher accuracy in predicting the right wrist. On the other hand, for the view (b) and view (c), where the joints of the human body are not severely occluded, the model assigns higher weights to each joint. Similarly, for joint 14 in the view (d), which corresponds to the left elbow, due to occlusion by the human body, the MRM model also assigns it a very small weight. Therefore, the MRM model's prediction of the left elbow is superior to the basic model.

Overall, when both the likelihood of MVJD module and the weight triangulation module are employed simultaneously, the model achieves the best predictive performance, with an MPJPE of 23.71. This result fully validates the synergistic effect between the likelihood of MVJD module

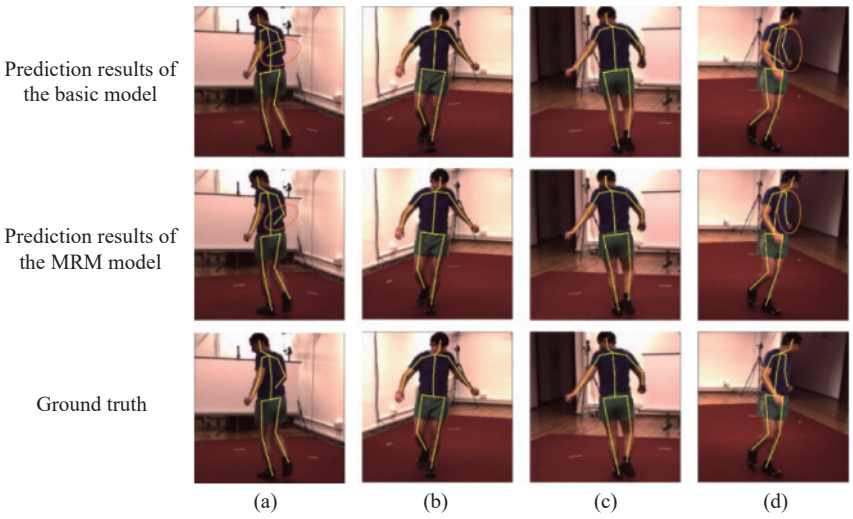


Fig. 3 Visual comparison of model predictions: (a) the first perspective; (b) the second perspective; (c) the third perspective; (d) the fourth perspective

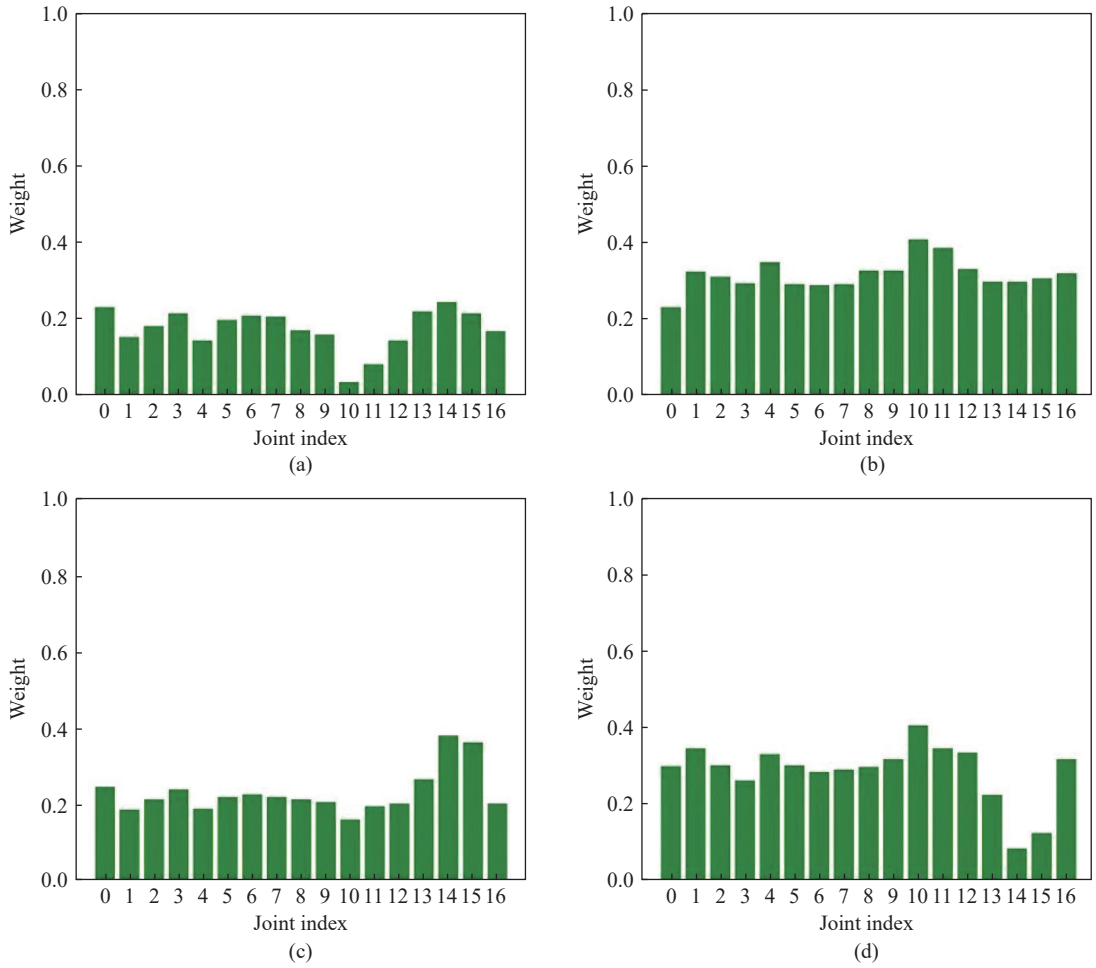


Fig. 4 Visual comparison of joint weights of MRM model: (a) the first perspective; (b) the second perspective; (c) the third perspective; (d) the fourth perspective

and the joint weighting mechanism. They implicitly and explicitly learn the connections among multiple viewpoints, respectively, complementing each other and jointly enhancing the model's capability to estimate 3D human poses from multiple viewpoints.

During the training of the MRM model, we conducted ablation experiments in both one-stage and two-stage approaches. The specific results are shown in Tab. 2. For the one-stage approach, we introduced a weight parameter λ and performed gradient backpropagation by adding the loss of likelihood of MVJD module L_{mle} and the loss of weight triangulation module L_{mse} together, i.e., $L_{\text{total}} = L_{\text{mle}} + \lambda L_{\text{mse}}$. For the two-stage approach, we first calculated the loss of likelihood of MVJD module L_{mle} and after minimizing L_{mle}, we froze the backbone network and the normalizing flow module of the model, only training the weight triangulation module to calculate L_{mse}. Initially, we believed that the optimization objectives of the two modules were significantly different, so we used a two-stage training approach. This method allowed the model to focus on different optimization objectives at each respective stage.

Tab. 2 Ablation results for one-stage and two-stage approach

Training approach	λ	MPJPE (mm)
One-stage	1.2	23.83
One-stage	0.8	**23.71**
One-stage	0.5	24.41
One-stage	–	23.95

However, the one-stage training approach can simplify the training process, and the experiment results are also quite impressive. By adjusting the training strategy of the weights, we bal-

anced the losses of the two modules during the model optimization, thus improving the overall prediction accuracy of the model. As shown in Tab. 2, by adopting a one-stage training approach with a weight of 0.8, the model's prediction accuracy has been improved. Overall, while the training approaches in the one-stage and two-stage do have a certain impact on model accuracy, the difference is not significant. This indicates that the likelihood of MVJD module and the weight triangulation module exhibit high flexibility and scalability.

4.5 Comparison Experiments

The MRM model was trained and validated on the Human3.6M dataset, with the results shown in Tab. 3. The MRM model demonstrates significant advantages over existing multi-view models across various actions, achieving superior results in 9 out of 14 action categories. Notably, in actions such as Dir., Pose and WalkD, the MRM model's MPJPE is significantly lower than that of other models, with reductions of 15.3%, 19.4%, and 5.2%, respectively, compared to the second-best model. This indicates that MRM possesses unique strengths in modeling the diversity of human motion. The average joint error of the MRM model is 23.7 mm, outperforming the TransFusion model [14] by 2.1 mm. Here, we are comparing with other heatmap-based models, rather than regression-based models. Heatmap-based models often outperform regression-based models in prediction accuracy, whereas the MRM model is based on regression. Nevertheless, the MRM model still achieved performance superior to that of heatmap-based models in terms of prediction accuracy. Moreover, models like Epipolar transformers [13] rely entirely on manually designed geometric constraints when aggregating multi-view features. This makes the models sensitive to camera parameters. In contrast, when aggregating multi-view features, MRM depends on learned complex multivariate joint distribution priors across different views, exhibiting flexibility with regard to camera parameters. Consequently, in experimental results, MRM demonstrates advantages across multiple actions. For actions like "sit" and "SitD.", although the MRM model has achieved a certain level of accuracy, its advantage is not pronounced compared to other models. We have noticed that the MRM demonstrates a significant improvement in prediction accuracy for complex actions like "sit" and "SitD." compared to its base model. The likelihood of MVJD module and the weight trian-

Tab. 3 Comparison with state-of-the-art methods on the Human3.6M dataset

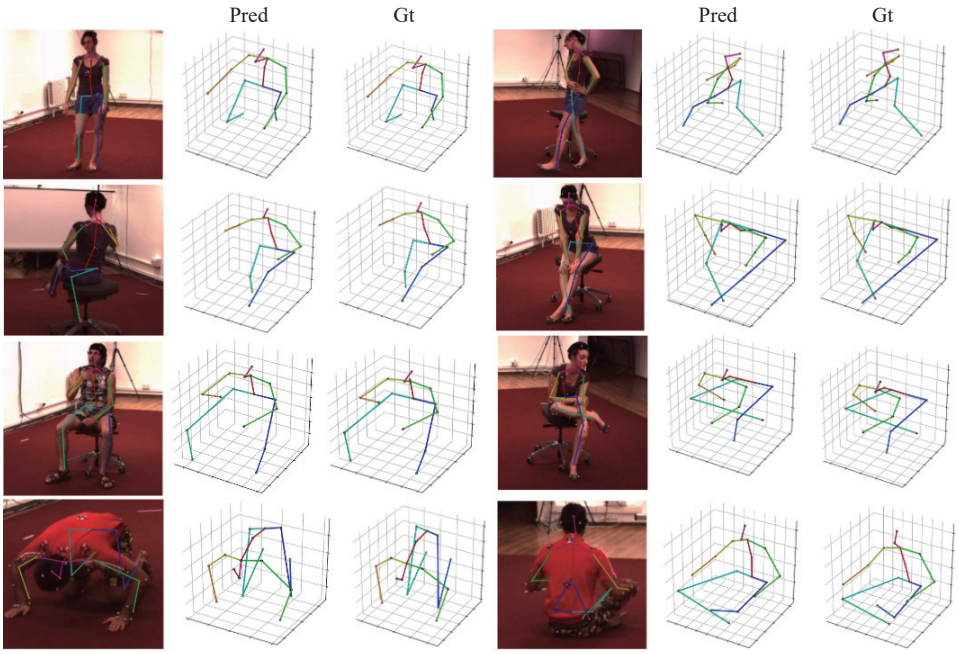
Method	Dir.	Disc.	Eat	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	walkT.	Average
Bartol et al. [29]	27.5	28.4	29.3	27.5	30.1	28.1	27.9	30.8	32.9	32.5	30.8	29.4	28.5	30.5	30.1	29.1
K.M. et al. [30]	39.4	46.9	41.0	42.7	53.6	54.8	41.4	50.0	59.9	78.8	49.8	46.2	51.1	40.5	41.0	49.1
He et al. [13]	25.7	27.7	23.7	24.8	26.9	31.4	24.9	26.5	**28.8**	31.7	28.2	26.4	23.6	28.3	23.5	26.9
Haoyu et al. [14]	24.4	26.4	23.4	21.1	25.2	26.8	23.2	24.7	33.8	29.8	26.4	24.2	23.2	26.1	23.3	25.8
Qiu et al. [1]	24.0	26.7	23.2	24.3	24.8	**22.8**	24.1	28.6	32.1	26.9	30.9	25.6	25.0	28.0	24.4	26.2
Edoardo et al. [16]	27.3	32.1	25.0	26.5	29.3	35.4	28.8	31.6	36.4	31.7	31.2	29.9	26.9	33.7	30.4	30.2
Shuai et al. [31]	24.6	28.3	26.0	26.5	30.0	32.1	26.0	26.3	33.5	39.7	28.4	26.4	31.1	**24.6**	24.9	28.5
Ma et al. [32]	21.8	26.5	**21.0**	22.4	23.7	26.7	23.1	23.2	27.9	30.7	24.6	23.3	21.2	25.3	22.6	24.4
Bultmann et al. [33]	22.4	24.0	22.2	21.7	24.0	23.9	22.1	22.6	26.0	**26.8**	24.5	22.8	24.6	21.8	21.3	**23.5**
MRM (without (w/o) m and w)	22.7	27.2	24.9	23.4	29.8	30.1	22.1	30.3	37.8	53.1	30.0	25.0	23.1	31.9	24.0	29.2
MRM (w/o m)	21.7	25.4	23.7	22.4	25.4	24.7	21.7	29.0	30.1	47.0	26.6	23.1	21.9	29.6	23.9	26.4
MRM (w/o w)	23.2	27.0	25.5	23.7	29.9	29.4	22.1	26.0	36.3	47.6	29.8	24.9	21.2	28.5	21.7	28.2
MRM	**19.5**	**22.8**	22.0	**20.1**	**23.4**	24.4	**20.0**	**21.8**	29.0	38.0	**24.0**	**22.2**	**20.2**	25.0	**21.1**	23.7

Note: All the values presented are MPJPE scores (mm). "m" and "w" refer to the multivariate joint distribution and joint weights respectively. Here Dir. represents directions, Disc. represents discussion, Purch. represents purchases, SitD. represents sitting down, WalkD. represents walk dog, walkT. represents walk together.

gulation module in the MRM actually play a role in predicting these complex actions. However, the bottleneck in its low prediction accuracy primarily lies in the poor feature extraction performance of the backbone network for the "sit" and "SitD." actions. On one hand, there are fewer instances of "sit" and "SitD." actions compared to general postures. On the other hand, these actions suffer from severe occlusion. Although multi-view image information is input, it remains challenging to recover complete information in extreme occlusion conditions.

Fig. 5 displays the ground truth and predicted values of some pose estimation samples, with "gt" and "pred" representing the ground truth and predicted values, respectively. It can be observed that there is significant occlusion in the second and fourth rows, especially for "sit" where the arms and legs are largely obscured. Relying solely on single-view data would result in the loss of crucial joint information. However, by fusing features from multiple views, the model effectively overcomes the limitations of single-view observations, enabling precise prediction of occluded joints. Even in cases of severe occlusion, MRM is still capable of predicting relatively accurate human poses. It is evident that the projected 2D coordinates, obtained by mapping the predicted 3D skeletal coordinates back onto the 2D plane, are quite accurate. This not only validates the model's accurate prediction capability in 3D space but also confirms its reliable projection performance on the 2D plane.

Tab. 4 presents a comparison of our method with other methods on the MPI_INF_3DHP dataset. MRM retains the backbone network pretrained on the Human3.6M dataset and retrains the likelihood of MVJD module as well as the weight triangulation module on the MPI_INF_3DHP dataset. From the experimental results, it can be observed that MRM possesses cross-dataset transfer learning capability. MRM surpassed Shin et al. [28] by 8.5 mm on MPJPE. Procrustes analysis is typically used to align predicted results with ground truth, mitigating errors caused by global rotation, translation, and scaling. However, the MRM model achieves a lower MPJPE even without this post-processing,



Note: The first column shows the human skeleton with predicted 3D coordinates projected onto a 2D plane, the second column represents the predicted results of human pose estimation, and the third column displays the ground truth values.

Fig. 5 Human pose estimation results

Tab. 4 Comparison of different models on the MPI_INF_3DHP dataset

Method	MPJPE (mm)	3DPCK (%)	AUC (%)
Chai et al. [34]	61.3	92.1	62.5
Zhang et al. [35]	54.9	94.4	66.5
Hu et al. [36]	42.5	97.9	69.5
Shin et al. [28] (PA)	50.2	97.4	65.5
MRM (w/o flow)	44.6	98.4	70.6
MRM (w/o weights)	47.6	97.0	68.8
MRM	**41.7**	**98.6**	**72.1**

Note: (PA) stands for Procrustes analysis performed on the prediction results. Better results are indicated by higher PCK/AUC scores and lower MPJPE values.

Tab. 5 Comparison of computational efficiency on Human3.6M

Method	Parameter (10^6)	MACs (10^9)	MPJPE (mm)
Davoodnia et al. [37]	65.4	208.7	26.4
He et al. [13]	68.1	406.5	26.9
Shuai et al. [31]	78.6	407.0	–
ResNet-152	60.0	136.4	–
Iskakov et al. (alg) [10]	+20.0	+73.4	**22.6**
MRM	+9.0	+5.6	23.7

Note: "+" indicates the remaining number of parameters and MACs after removing the backbone network from the model.

suggesting that its predictions are closer to the ground truth.

4.6 Running Time Analysis

Tab. 5 compares the computational time and complexity, with the number of input views set to 4. It can be observed that, compared to the heatmap-based model [10], the MPJPE increases by only 7%, while the numbers of parameters and multiply-accumulate operations (MACs) decrease by 13.8% and 32.3%, respectively. For a more intuitive comparison, Tab. 5 shows the number of parameters and MACs of the backbone network ResNet-152. Since the likelihood of MVJD module in the MRM model is not required during inference, it does not participate in computations at that stage. Excluding the computations of the backbone network, the MRM model needs only 7.6% of the multiply-add operations compared to methods based on heatmaps.

4.7 Number of Views

Fig. 6 shows the MPJPE with different numbers of views on Human3.6M. It can be seen that when there are only two views, the MRM model achieves good results with the help of the likelihood of MVJD module. However, as the number of views increases, the joint weights can better integrate information from different views, enhancing the model's performance. Therefore, the MRM model has a certain stability regarding the change in the number of views, and it does not lose accuracy even when the number of views is low.

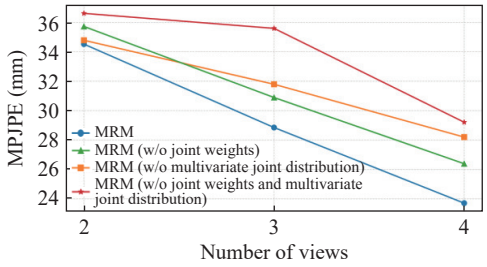


Fig. 6 MPJPE with different numbers of views

5 Conclusion

This paper delves deeply into the problem of multi-view 3D human pose estimation based on multivariate joint distribution regression and proposes an innovative multi-view 3D human pose estimation model, MRM. The model aims to predict 3D joint coordinates through direct regression while fully leveraging the rich and complementary information present in multi-view data to enhance the accuracy and robustness of human pose estimation. Compared to heatmap-based models, the MRM model can significantly reduce computational resources while maintaining high accuracy. To our knowledge, we are the first to propose learning the multivariate joint distribution of multi-view human poses through a normalizing flow model. This approach enables the model to learn the relationships among human poses from different viewpoints and provides valuable prior knowledge during model inference. The MRM model comprises three main components: the backbone network, the likeli-

hood of MVJD module, and the weight triangulation module. The backbone network is responsible for extracting image features and joint weights from the input multi-view images, providing a foundation for subsequent processing. The likelihood of MVJD module captures the inherent connections among different viewpoints by learning the joint distribution of multi-view pose data, thereby enhancing the model's capability to handle occlusion scenarios. By leveraging joint weights to learn the differences among various views and jointly training images from different views, the model can improve its robustness to occlusion during triangulation. The weight triangulation module combines the joint weights extracted from the backbone network to accurately infer 3D joint coordinates from 2D coordinates using linear algebraic triangulation methods. Experimental results indicate that the MRM model achieves significantly superior performance compared to existing methods on two widely used benchmark datasets for 3D human pose estimation, namely Human3.6M and MPI_INF_3DHP. In particular, when dealing with occlusion scenarios, the MRM model demonstrates strong robustness and accuracy. Furthermore, as a regression-based approach, MRM can significantly reduce the number of MACs required by the model, which is crucial for edge device applications. Meanwhile, the MRM model maintains stable performance across varying numbers of viewpoints, indicating its good transferability and practical application potential.

However, there is still room for improvement in the MRM model. When predicting complex actions like SitD. in the Human3.6M dataset, the error fluctuations are relatively large compared to other actions. In the subsequent steps, we can attempt to generate more training data samples of complex actions through data augmentation techniques or generative adversarial networks (GANs), so that the model can learn the pose distributions of a wider variety of complex actions. Additionally, we will introduce an attention mechanism into the model to enable it to focus more on the key joints and pose variations in complex actions.

References:

[1] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross view fusion for 3D human pose estimation, " in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, South Korea, pp. 4341-4350, 2019.

[2] J. Li, S. Bian, A. Zeng, C. Wang, B. Pang, W. Liu, and C. Lu, "Human pose regression with residual log-likelihood estimation, " in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), Montreal, QC, Canada, pp. 11005-11014, 2021.

[3] Z. Wang, X. Nie, X. Qu, Y. Chen, and S. Liu, "Distribution-aware single-stage models for multi-person 3D pose estimation, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, pp. 13086-13095, 2022.

[4] M. Saied, F. Adjogble, S. Guirguis, M. Hemmji, and J. Warschat, "A framework for systematic scientific research management, " in *Portland International Conference on Management of Engineering and Technology* (*PICMET*), Monterrey, Mexico, pp. 1-16, 2023.

[5] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP, " in *Proceedings of the 5th International Conference on Learning Representations* (*ICLR*), Toulon, France, pp. 1-32, 2017.

[6] D. Jimenez Rezende and S. Mohamed, "Variational inference with normalizing flows, " *ArXiv Preprint*, arXiv: 1505.05770, 2015.

[7] L. Dinh, D. Krueger, and Y. Bengio, "Nice: Nonlinear independent components estimation, " *ArXiv Preprint*, arXiv: 1410.8516, 2014.

[8] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu,"Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325-1339, 2014.

[9] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, and W. Xu, "Monocular 3D human pose estimation in the wild using improved CNN supervision, " in *International Conference on 3D Vision* (*3DV*), Qingdao, China, pp. 506-516, 2017.

[10] K. Iskakov, E. Burkov, V. Lempitsky, and Y.

Malkov, "Learnable triangulation of human pose, " in *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, South Korea, pp. 7717-7726, 2019.

[11] H. Tu, C. Wang, and W. Zeng, "VoxelPose: Towards multi-camera 3D human pose estimation in wild environment, " in *Computer Vision – ECCV 2020: 16th European Conference, August 23–28, 2020, Proceedings, Part I*, Glasgow, UK, pp. 197-212, 2020.

[12] H. Ye, W. Zhu, C. Wang, R. Wu, and Y. Wang, "Faster VoxelPose: Real-time 3D human pose estimation by orthographic projection, " in *Proceedings of the European Conference on Computer Vision* (*ECCV*), Tel Aviv, Israel, pp. 142-159, 2022.

[13] Y. He, R. Yan, K. Fragkiadaki, and S. I. Yu, "Epipolar transformers, " *ArXiv Preprint*, arXiv: 2005.04551, 2020.

[14] H. Ma, L. Chen, D. Kong, Z. Wang, X. Liu, H. Tang, X. Yan, Y. Xie, S. Y. Lin, and X. Xie, "TransFusion: Cross-view fusion with transformer for 3D human pose estimation, " *ArXiv Preprint*, arXiv: 2110.09554, 2021.

[15] S. Chun, S. Park, and J. Y. Chang, "Representation learning of vertex heatmaps for 3D human mesh reconstruction from multi-view images, " in *2023 IEEE International Conference on Image Processing* (*ICIP*), Kuala Lumpur, Malaysia, pp. 670-674, 2023.

[16] E. Remelli, S. Han, S. Honari, P. Fua, and R. Wang, "Lightweight multi-view 3D pose estimation through camera-disentangled representation, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, pp. 6039-6048, 2020.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need, " in *Advances in Neural Information Processing Systems* (*NIPS*), Long Beach, CA, USA, pp. 5998-6008, 2017.

[18] Z. Zhang, C. Wang, W. Qiu, W. Qin, and W. Zeng,"AdaFuse: Adaptive multiview fusion for accurate human pose estimation in the wild," *International Journal of Computer Vision*, vol. 129, no. 3, pp. 703-718, 2021.

[19] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback, " in *2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, pp. 4733-4742, 2016.

[20] S. Liang, X. Sun, and Y. Wei,"Compositional human pose regression," *Computer Vision and Image Understanding*, vol. 176, no. 1, pp. 1-8, 2018.

[21] J. Martinez, R. Hossain, J. Romero, and J. Little, "A simple yet effective baseline for 3D human pose estimation, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, pp. 2640-2649, 2017.

[22] W. Shan, Z. Liu, X. Zhang, Z. Wang, K. Han, S. Wang, S. Ma, and W. Gao, "Diffusion-based 3D human pose estimation with multi-hypothesis aggregation, " in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, France, pp. 14761-14771, 2023.

[23] W. Zhu, X. Ma, Z. Liu, L. Liu, W. Wu, and Y. Wang, "MotionBERT: A unified perspective on learning human motion representations, " in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (*ICCV*), Paris, France, pp. 15085-15099, 2023.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, NV, USA, pp. 770-778, 2016.

[25] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context, " in *Computer Vision – ECCV 2014: 13th European Conference, September 6–12, 2014, Proceedings, Part V*, Zurich, Switzerland, pp. 740-755, 2014.

[26] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis, " in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, pp. 3686-3693, 2014.

[27] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei, "Integral human pose regression, " *ArXiv Preprint*, arXiv: 1811.08236, 2018.

[28] S. Shin and E. Halilaj, "Multi-view human pose and shape estimation using learnable volumetric aggregation, " *ArXiv Preprint*, arXiv: 2011.13427, 2020.

[29] K. Bartol, D. Bojanic, T. Petkovic, and T. Pribanic, "Generalizable human pose triangulation, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), New Orleans, LA, USA, pp. 11018-11027, 2022.

[30] A. Kadkhodamohammadi and N. Padoy,"A gener-

alizable approach for multi-view 3D human pose regression," *Machine Vision and Applications*, vol. 32, no. 1, pp. 1-14, 2021.

[31] H. Shuai, L. Wu, and Q. Liu,"Adaptive multi-view and temporal fusing transformer for 3D human pose estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4122-4135, 2023.

[32] H. Ma, Z. Wang, Y. Chen, D. Kong, L. Chen, X. Liu, X. Yan, H. Tang, and X. Xie, "PPT: Token-pruned pose transformer for monocular and multi-view human pose estimation, " in *Computer Vision – ECCV 2022: 17th European Conference, October 23–27, 2022, Proceedings, Part V*, Tel Aviv, Israel, pp. 424-442, 2022.

[33] S. Bultmann and S. Behnke, "Real-time multi-view 3D human pose estimation using semantic feedback to smart edge sensors, " *ArXiv Preprint*, arXiv: 2106.14729, 2021.

[34] W. Chai, Z. Jiang, J. N. Hwang, and G. Wang, "Global adaptation meets local generalization: Unsupervised domain adaptation for 3D human pose estimation, " *ArXiv Preprint*, arXiv: 2303.16456, 2023.

[35] J. Zhang, Z. Tu, J. Yang, Y. Chen, and J. Yuan, "MixSTE: Seq2seq mixed spatio-temporal encoder for 3D human pose estimation in video, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 13222-13232, 2022.

[36] W. Hu, C. Zhang, F. Zhan, L. Zhang, and T. T. Wong, "Conditional directed graph convolution for 3D human pose estimation, " in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 11417-11426, 2021.

[37] V. Davoodnia, S. Ghorbani, M. A. Carbonneau, A. Messier, and A. Etemad, "UPose3D: Uncertainty-aware 3D human pose estimation with cross-view and temporal cues, " in *Computer Vision – ECCV 2024: 18th European Conference, September 29–October 4, 2024, Proceedings, Part VII*, Milan, Italy, pp. 19-38, 2024.

Anzhan Liu is presently an Associate Professor in Zhongyuan University of Technology, Zhengzhou, China. He received the Master's degree in computer software and theory from Huazhong University of Science and Technology, Wuhan, China. in 2008. His research focuses on the computer vision, machine and deep learning, intelligent computing, etc.

Hufei Zhao received the B.S. degree from Yancheng Institute of Technology, Yancheng, China in 2022 and the M.S. degree from Zhongyuan University of Technology, Zhengzhou, China in 2025. His research interests include human pose estimation and object detection

Yilu Ding received the B.E. degree from North China University of Water Resources and Electric Power of Computer Science and Technology, Zhengzhou, China, in 2022. She is currently studying toward the M.S. degree in Zhongyuan University of Technology, Zhengzhou, China. Her research interests include computer vision and human pose estimation.