A Multiview Approach for Pedestrian 3D Pose Detection and Reconstruction

Kai Chen[1], Xiaodong Zhao[1(✉)], Yujie Huang[1], and Pengfei Wang[2]

[1] Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
xdzhao@nuaa.edu.cn

[2] Nanjing Research Institute of Electronic Engineering, Nanjing 210007, China

Abstract. Aiming at the problems that most of the existing methods of constructing 3D models of human body based on 2D human body surface pose points will lead to continuous modeling jitter and local distortion of the modeling results, we propose a 3D pose point detection method based on the skinned multiplayer linear model (SMPL) in the human body, which maps the 2D pose points of the body to 3D pose points of the real scene in the multiview perspective through a clustering algorithm, and introduces Kalman filtering to de-noise the human body pose points. The Kalman filter is introduced to denoise the human body posture points. In the process of constructing a 3D model of the human body based on 3D pose points, we construct an end-to-end human 3D modeling network (SMPL-VAE) based on the correction of gradient descent regression network by the automatic variational approach (VAE), which is more in line with the local modeling of the human body's motion structure while maintaining the overall proportion.The results on open dataset Shelf show that our methods improve the quality of human post point detection and modeling.

Keywords: Multi-Ocular Vision · Skinned Multi-Person Linear Modeling (SMPL) · 3D Pose Detection

1 Introduction

With the continuous development of science and technology and the significant improvement of urban informatization level, future robots and intelligent systems in the real world need to be able to perceive and understand human beings from visual inputs and interact with the real world, such as digital twin modeling for intelligent factories and panoramic imaging system for in-vehicle AVMs, and other technologies. Therefore, 3D modeling of human body in three-dimensional space is of great significance.

Multi-target detection and 3D modeling generally refer to the detection of pedestrians under the viewpoint of the device when the number of unknown targets is unknown, and modeling in 3D space in order to carry out the subsequent trajectory prediction and analysis. In practical application scenarios, multi-target detection not only needs to be real-time, but also needs to be accomplished between different targets in each frame, which puts forward higher requirements for detection and modeling due to the respective

S. Saito et al. (Eds.): AsiaSim 2024, CCIS 2170, pp. 87–100, 2024.
https://doi.org/10.1007/978-981-97-7225-4_7

differences between the targets as well as the intricate correlation situation between the pose points. Most of the existing multi-target detection algorithms [1–4] start from the surface of human body, and mainly improve on accuracy, computational resources and algorithms [5–7]. However, they are unable to solve the problems of target occlusion and target matching with pose points well. In terms of modeling, although the existing methods [8–10] have achieved better modeling results in terms of reconstruction error, floating point operations per second (FLOPs), etc., they are still prone to local pose distortion and jitter in continuous modeling when there is more occlusion and the target cascade relationship is more complex.

Aiming at the existing problem of detecting and modeling the 2D body surface pose points of dense pedestrians under binocular vision, we propose a 3D pose point detection method inside the human body based on the Skinned Multi-Person Linear Model [11] (SMPL), and then construct an end-to-end SMPL-VAE human body regression model by noise reduction and smoothing of 3D pose points inside the body through Kalman Filtering: the step-by-step training and the joint training are combined, and the VAE training results are used to correct the modeling results of the SMPL, to take into account the overall proportion and at the same time to ensure that the local modeling results conform to the structure of the human body movement more closely.

2 Related Works

Thanks to the wide application of deep learning technology, the field of 2D human pose detection has made remarkable development. DeepPose [4] extends the AlexNet [14] network, adopts DNN [15] to learn the human pose point feature representations, and optimizes the results of the previous processing step through a cascade structure; OpenPose proposes a PAF+ confidence-associated detection network that performs feature extraction on a single RGB image, which effectively solves problems such as low accuracy of key human pose detection due to partial occlusion. However, associating only the surface pose of 2D images cannot accurately reflect the complex body poses of the human body, and the final 3D skeleton projection and body modeling will lose the stereoscopic sense because of neglecting the thickness of the human body itself; moreover, with the increase in the number of targets, the detection performance and efficiency of 2D pose detection methods are poor in dealing with problems such as multi-person occlusion.

Based on the inadequacy of 2D human pose detection techniques and the lack of depth information, 3D human pose detection techniques [16–18] were developed.Tekin [16] et al. established a deep learning regression framework by fusing convolutional neural network CNN with recurrent neural network RNN; Moon G et al. [17] proposed a camera distance sensing framework to extract the target from a 2D image, using RootNet to localize the root node of the human body and PoseNet to localize the relative root node of the human body to maintain spatial consistency and obtain 3D human pose detection; VoxelPose [18] operates directly in the 3D space, avoiding the case of misdetection for each camera viewpoint. Although these methods achieve better detection results, they still cannot effectively solve the problems of target occlusion as well as correlation and matching between pose points when facing dense scenes.

As 3D modeling of the human body continues to emerge as a promising application in various fields, this technology is gradually attracting scholars to invest in research: SMPLify [19] uses a CNN network to estimate the posture points on the 2D body surface, and then uses a statistical model to fit them to the 2D joints; HMR, an end-to-end human posture detection network proposed by Kanazawa et al. [20] adopts a regressive approach, where images are processed through an encoder and then fed into a 3D regression module as well as a discriminator to minimize the reprojection error before determining the authenticity of the human pose; MVPOSE [21] introduces a recursive Bayesian filter function to detect each pedestrian individually as a way of reducing the problem of correlation determination between the pedestrian's 3D pose points. Although these methods have also achieved fruitful results, when facing the situation of human detection modeling in a huge state space, due to the poor treatment of local poses of individuals in a dense state, there are still obvious distorted modeling results in some wrist joints and other places.

3 Proposed Method

To address the problems of mis-correlation and matching of pose points, poor local poses, and jitter in continuous modeling results caused by the complex cascading relationship between dense pedestrians occluding each other under binocular vision in related work. In this section, the method we proposed in this will be introduced in detail, which mainly focuses on two major aspects of 3D intra-personal pose detection as well as modeling.

3.1 Multi-target Intra-personal 3D Pose Detection

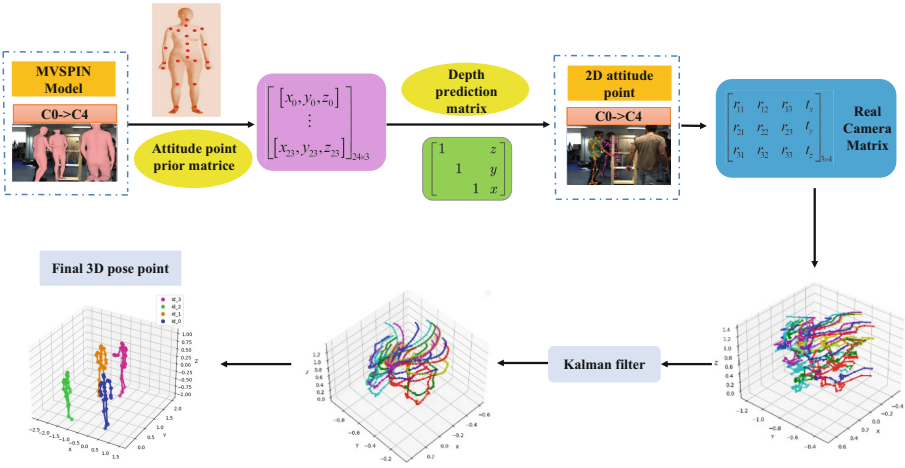


Fig. 1. 3D pose point detection framework.

In the actual multi-view pedestrian detection applications, there are differences in the parameter models, shooting angles, and viewing angles of different cameras, as well as

possible occlusion problems between different viewing angles, so it is difficult to obtain accurate pose point data, which increases the complexity of tasks such as pedestrian trajectory prediction and 3D modeling. Therefore, it is particularly important to further improve the human posture detection methods to realize more realistic and effective multi-view pedestrian association.

As shown in Fig. 1, we propose a 3D pose point detection method based on the Skinned Multi-Person Linear Model (SMPL) for the interior of the human body, which obtains the SMPL rendering model of a single view angle and the joint pose a priori matrices through the MVSPIN model, and then computes the 3D predicted pose points that contain the depth prediction information. In order to integrate the information of each viewpoint and enhance the correlation, this paper adopts the camera matrix generated by the weak perspective model to map the 3D predicted pose points, and obtains the 2D pose points under the corresponding viewpoints; and then adopts the real camera matrix to correlate the 2D pose points of each viewpoint, and returns to the 3D space; at the same time, a Kalman filter is introduced to smooth the human pose points and solve the jitter problem of continuous modeling, and finally realizes the multiview multiple pose points under the spatial scene.

2D Image Pose Point Detection Based on MVSPIN. We use the MVSPIN pose detection model to obtain the rendering model information of each target in the planar im-age, and extract 6890 surface pose point coordinates, while obtaining the joint regres-sion prior matrices from the SMPL basic model, which in turn computes the coordina-tes of 24 3D predicted pose points in the generalized human body model.

$$pose\_pre\_result = J\_regressor\_prior \times v\_posed \tag{1}$$

where $J\_regressor\_prior$ is the joint regression prior matrix, v_{posed} is a matrix containing 24 3D pose points.

Since the MVSPIN model is based on a single view for 3D pose point detection, the limitations that are present in the single view angle make errors in the longitudinal depth direction. In addition, due to camera angle, occlusion, etc., there is also a deviation between the images from different viewpoints for the predicted 3D pose and the original image. To solve this problem, this paper considers eliminating the depth data predicted for each view and restoring the 3D coordinates to the 2D plane of the original image based on the depth prediction matrix generated during the construction of each target model. In this way, preparation is made for a more accurate restoration of the 3D information on the image.

As shown in Fig. 2, (a) the camera prediction parameters are transformed into a 3 $\times$ 4 camera prediction matrix by a unit matrix; (b) the 3D pose points predicted by the SMPL model are transformed into a 1 $\times$ 24 $\times$ 3 $\times$ 1 pose matrix; and (c) the two matrices are multiplied together to obtain the coordinates of the planar 2D pose point 1 $\times$ 24 $\times$ 2 $\times$ 1. In this paper, the predicted 3D pose points are matrix transformed by constructing a 4 $\times$ 4 unit matrix and multiplied with the camera prediction matrix to obtain the coordinates of the pose points in the camera direction of the image prediction.In order to eliminate the error of depth prediction, we eliminate the depth component in the camera direction to obtain the 2D pose point coordinates. This method ensures the completeness of individual target pose points and takes the thickness of pedestrians into account, which provides a good foundation for the following regression from 2D to 3D poses.

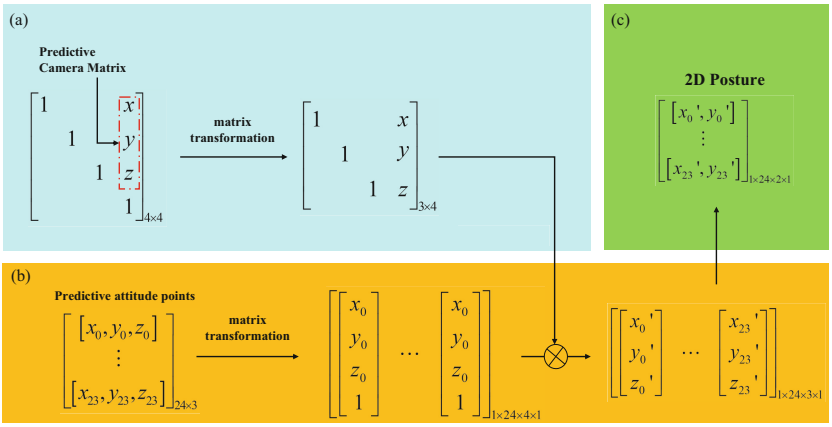


Fig. 2. 2D pose point restoration under various viewing.

3D Pose Point Regression and Filtering Processing. We use the real camera matrix to reproject the 2D pose points of all views into 3D spatial.Since 3D pose point detec-tion is performed in multiple views, there is no need to obtain depth data. It is suffici-ent to regress the 3D coordinates through the vertically projected 2D coordinates of each view and the real camera matrix.

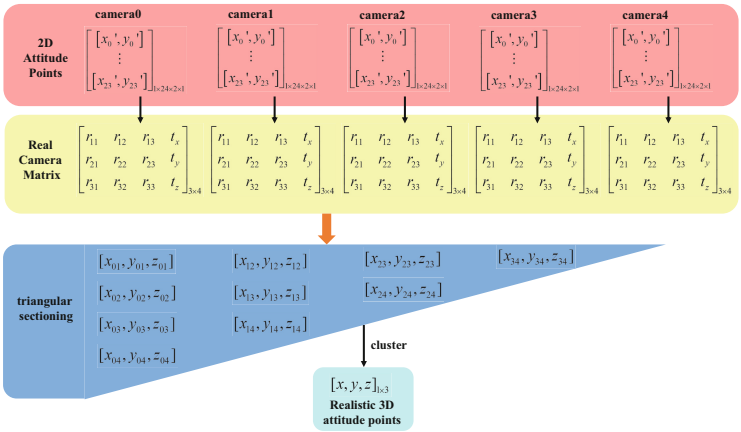


Fig. 3. 3D coordinate regression.

As in Fig. 3, based on the 2D pose point coordinates calculated above, each view is processed with a real camera matrix, and according to the monoclinic constraints existing between the planes [22], the matching and reconstruction between the corresponding points of the two planes are realized by triangular dissection [23] to obtain the 3D pose point coordinates containing the thickness of the human body, in order to better characterize the real position of the pedestrian in space.

Since the changes of spatial targets in the application scenario are real-time dynamic, the uncertainty of local body parts is high, and the flexibility is strong, some of the poses may have short-time large changes, which will lead to the human body modeling error becomes larger and damage the effect od modeling. Therefore, we regarded each pose point (*x, y, z*) of the human body as a state, and Kalman filtering is used to estimate the changes of these states.

Human actions have autonomous consciousness and are not subject to external intervention and control, we introduce the control input **B** and the process noise covariance **Q** in the prediction and update them with the initial form of the zero matrix of 3×3, while the observed state is directly measured by the measurement matrix **H** without passing through a complex nonlinear transformation. Therefore **H** is also initialized to the unit matrix of 3×3, the state vector is initialized with the initial observations, and the state covariance matrix is set to a larger value to reflect the uncertainty about the initial state.

The whole filtering process is carried out in two major steps, prediction and update [21], and the system model is used to predict the state $\widehat{X}_k$ at the next time step as in Eq. 2. The state covariance $\widehat{P}_k$ is also predicted to determine the uncertainty of the prediction as in Eq. 3.

$$\widehat{X}_k = F \times \widehat{X}_{k-1} + B \times u_k \tag{2}$$

$$P_k = F \times P_{k-1} \times F^T + Q \tag{3}$$

where $\boldsymbol{F}$ is the state transfer matrix, $\boldsymbol{B}$ is the control input matrix, u_k is the control input, for which it is initialized as a one-dimensional row vector of length 3.

The update is also performed on the state and covariance, as in Eqs. 5 and 6 to correct the predicted state estimate through the measurements z_k as well as the Kalman gain K_k, and then the covariance is updated to keep track of the uncertainty in the state estimate of the system, and to provide an optimal state estimate in the presence of uncertainty.

$$K_k = \frac{P_k \times H^T}{H \times P_k \times F^T + R} \tag{4}$$

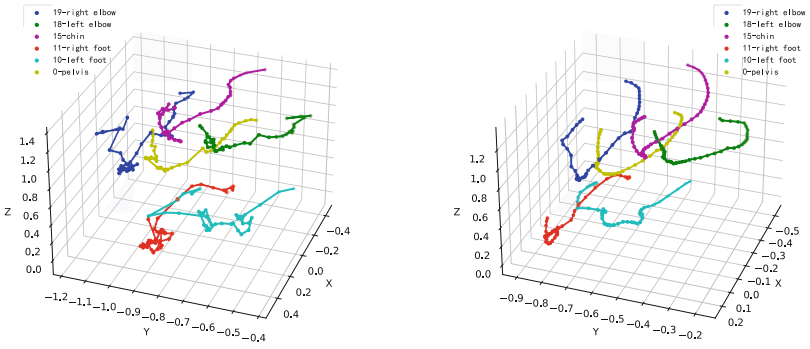
$$\widehat{X}_k = \widehat{X}_k + K_k \times \left(z_k - H \times \widehat{X}_k\right) \tag{5}$$

$$P_k = (I - K_k \times H) \times P_k \tag{6}$$

where $\boldsymbol{I}$ is the unit matrix, $\boldsymbol{R}$ is the measurement noise covariance matrix.

As shown in Fig. 4 for the comparison of Kalman filtering effect, we select 0-pelvis, 10-left foot, 11-right foot, 15-chin, 18-left elbow and 19-right elbow with a larger range of activity for comparison of the pose points, which can be seen that the filtered trajectory of the movement is smoother.

After the acquisition of 2D pose points from each viewpoint and 3D pose regression, and then Kalman filtering, the 3D human pose points with good poses and smooth continuous modeling are finally obtained, which lays a good foundation for 3D modeling of the human body.



(a) Without Kalman filtering

(b) With Kalman filtering

Fig. 4. Comparison of Kalman filter processing effect

3.2 3D Modeling of Human Body Based on SMPL-VAE

Since the work in the section of 3.1 is global processing for each target as a whole and no local processing is applied, if these pose points are directly used for 3D modeling of the human body, the modeling of local poses of human joints will be distorted.

Therefore, for the problem of 3D modeling of the human body with processed pose points, we refer to the idea of VAE [13] and propose the SMPL-VAE model, as shown in Fig. 6, which combines step-by-step training with joint training to generate 3D modeling results of the human body. The two-way step-by-step training processes the input pose points in different ways, and then the final modeling results are obtained by joint training.

Step-by-step training I: The 3D pose point X finally obtained in the section of 3.1 is taken as input, and the SMPL parameter $\Theta_S(\theta, \beta)$ is trained by the iterative optimization method of gradient descent [24] using the ground truth as a criterion, and the data is continuously fitted to make it closer to the ground truth. Due to the high degree of autonomy and irregularity of the target activity, a loss function that is robust to outliers is needed, so we choose to use Smooth L1 Loss [25], whose strong robustness provides more stable convergence and more reliable error gradients than other loss functions. This loss function strikes a balance between the mean square error and the absolute error, thus reducing the impact of large error values and ensuring a smoother and more gradual optimization process as in Eq. 7.

$$L_{Smpl} = \begin{cases} \frac{1}{2}\left(X_{gt} - J\right)^2, if \left|X_{gt} - J\right| < 1 \\ \left|X_{gt} - J\right| - \frac{1}{2}, if \left|X_{gt} - J\right| \geq 1 \end{cases} \tag{7}$$

where $\boldsymbol{J}$ is the coordinates of the human joint points output by the model and X_{gt} is the ground truth of the human joint points (Fig. 5).

In this way, we are able to obtain the human body's mesh vertices *verts* and human body joints coordinates J, where *verts* size is $n \times 6890 \times 3$, n is the number of targets in the scene during modeling, and $\boldsymbol{J}$ size is $n \times 24 \times 3$. The essence of this method lies in its ability to refine the model parameters during iteration, and more closely and accurately represent the human body pose 3D step-by-step ground-truthing values. Up

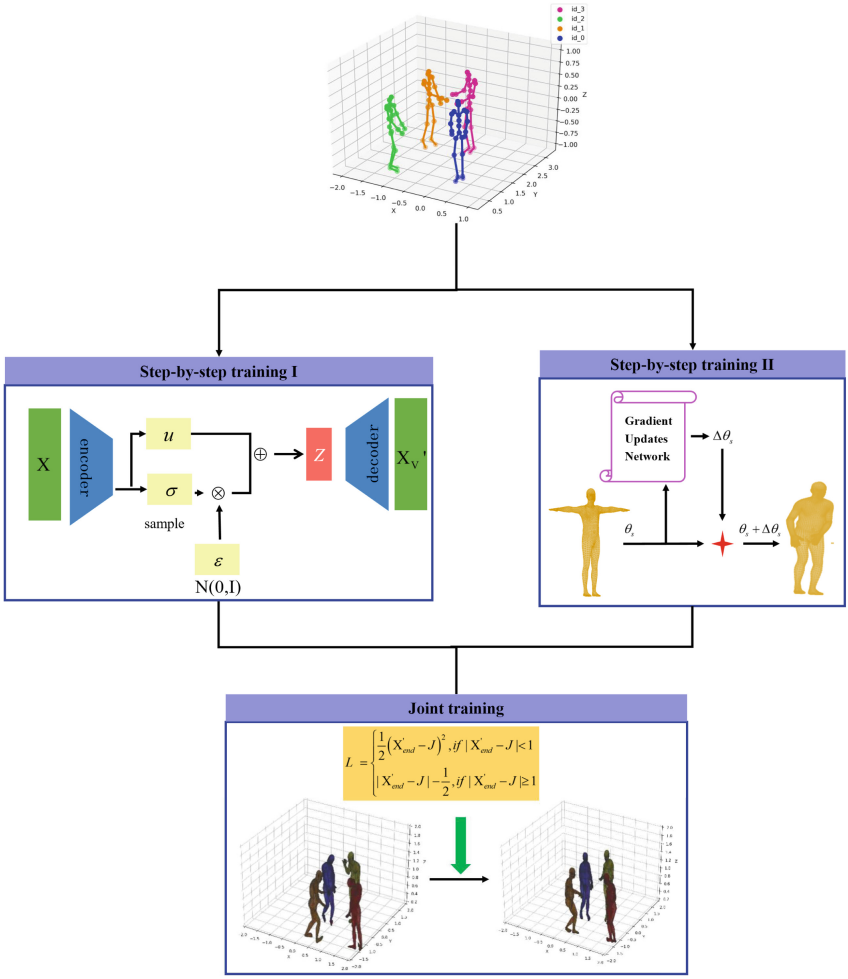


Fig. 5. Human body 3D modeling framework

to this point, step-by-step training one obtains preliminary human 3D modeling results, but in the local joints of the human body, there is still a violation of the laws of human kinematics, which still does not meet the final modeling needs.

Step-by-step training II: We use a Variable Auto-Encoder (VAE) framework to encode and decode the filtered dataset.The VAE obtains the generated 3D human pose points $\boldsymbol{X_v}$ by learning a latent space representation of the input data $\boldsymbol{X}$, and then reconstructing the inputs using this latent space.This process introduces a relaxation mechanism designed to learn the substeps on the latent space, allowing for generation of new samples, thus accommodating human pose transformations that conform to kinematic constraints.

The encoding phase plays an important role by compressing $\boldsymbol{X}$ into a lower dimensional space that captures the stepwise characteristics of the potential variables needed

for reconstruction, and the decoding phase attempts to reconstruct the input data from this compressed representation, preserving the critical structure without compromising the global and local fidelity of the 3D modeling of the human body, as in Eqs. 8 and 9 respectively.

$$z \sim q(z|\boldsymbol{X}) \tag{8}$$

$$\boldsymbol{X}'_V \sim p(\boldsymbol{X}|z) \tag{9}$$

where $q(z|\boldsymbol{X})$ is the latent variable stepping defined by the encoder; $p(\boldsymbol{X}|z)$ is the conditional stepping defined by the generator.

In the training process, then $\Theta_V(\phi_{en}, \phi_{de})$ is trained, which is done by maximizing the likelihood of the data. However, solving the likelihood function directly is difficult because the latent variables in VAE are continuous. To solve this problem, VAE uses a regularization term that learns the distribution of latent variables to simplify solving the likelihood function. The loss function consists of two parts: reconstruction loss and regularization loss. The reconstruction loss is the difference between the output sample and the original sample in VAE when the input sample is transformed into latent variables by the encoder and then the output sample is generated by the decoder. The smaller the reconstruction loss, the better the performance of the generator. Regularization loss, on the other hand, refers to the difference between the substeps of the latent variables and the a priori substeps in VAE by introducing the KL divergence [26]. This difference is quantified by the KL divergence, with a smaller KL divergence indicating that the latent variable is closer to the a priori substep. Taken together, the loss function of the VAE can be tabulated as:

$$L_{Vae} = -\mathrm{E}\big[\log p(X|z)\big] + \alpha \cdot KL(q(z|X)||p(z)) \tag{10}$$

where α is the tradeoff of the parameters, and $p(z)$ is the prior step. Finally, the results of step-by-step training I and step-by-step training II are fused for joint training. Using the results of step-by-step training I as a starting point, the output of step-by-step training I is corrected according to the results of step-by-step training II to better converge to the actual situation. And the trained parameters Θ (Θ_S, Θ_V) in the step-by-step training are still used for the second training using Smooth L1 Loss, as in Eq. 11, to obtain the final trimmed human pose points X'_{end}.

$$L = \begin{cases} \frac{1}{2}\left(X'_{end} - J\right)^2, if \left|X'_{end} - J\right| < 1 \\ \left|X'_{end} - J\right| - \frac{1}{2}, if \left|X'_{end} - J\right| \geq 1 \end{cases} \tag{11}$$

where X'_{end} is the final human joint point coordinates obtained after joint training.

This joint training model allows for knowledge transfer between the two methods. The iterative refinement of the parameters is enhanced by the relaxed constraints imposed by the VAE, and the excellent results of the overall model benefit from the ability to adjust local modeling effects. This combined training approach not only improves the accuracy of the pose estimation, but also ensures that the model adapts to the diversity of human movements, contributing to a more natural and intuitive interaction between humans and the digital environment.

4 Experiments

4.1 Introduction to Public Datasets and Evaluation Metrics

The Shelf [27] dataset was used in this study as a benchmark for model testing. The dataset contains an indoor scene with four people carrying shelves in close proximity and is equipped with five calibrated cameras. Each view presents a different degree of occlusion and the pedestrians may appear to be added, completely disappeared, reappeared, and other complexities, which provides favorable conditions for algorithm testing. The dataset provides a more realistic representation of the pedestrians' action postures, taking into account factors that may have an impact on the tasks of detection, tracking and feature association.

In the process of human 3D pose detection and modeling, we take the ground truth of the complete 3D pedestrian pose point coordinates as the standard in the training process, and due to the problems of missing targets and ID errors in the real values provided by Shelf, we re-labeled 1500 images of the dataset in 5 views based on the 24 bit-pose points of the SMPL model.

Considering that the mean positional error per joint [28] (MPJPE) is expressed in terms of the physical positional errors of the joints, which makes the results more intuitive and easy to understand, the error per joint can help to analyze the model's performance in different parts, thus providing more detailed feedback. Therefore, the quantitative evaluation metric we take is MPJPE, the smaller its value, the better the performance of the model in estimating the joint positions, as in Eq. 12.

$$MPJPE = \frac{1}{N}\sum_{i=1}^{N} \|p_i - \hat{p}_i\| \tag{12}$$

where $\|\cdot\|$ is the Euclidean distance, i.e., the distance between the true joint position and the model-predicted joint position, N is the 24 joint points of the human body, p_i is the true position of each joint point; and $\hat{p}_i$ is the position predicted by the model.

Percentage of Correct Keypoints [29] (PCK) focuses on the location of keypoints, provides an intuitive percentage representation of keypoint accuracy, applies to different thresholds, is flexible and versatile, and facilitates the comparison of the performance of different models on the same task, so we select PCK as an evaluation metric as in Eq. 13.

$$PCK = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\|p_i - \hat{p}_i\|}{D} \leq T\right) \times 100\% \tag{13}$$

where T is the threshold value, which is selected as 160mm,D is the reference length, which is set as the normalization factor of a single target, δ is the indicator function, which takes the value of 1 if the condition is true and 0 otherwise.

As shown in Table 1, the method proposed in this paper has demonstrated ideal test results. It improved by 3.88 on the MPJPE metric compared to the well-performing Shape-Aware, and it improved by 3.50 on the PCK metric compared to the well-performing SMPLify.

Table 1. Comparison with previous work based on Shelf dataset.

method	MPJPE	PCK
SMPL	107.12	86.24
SMPLify	109.30	88.76
HMR	114.62	83.26
Shape-aware	105.62	85.35
our	101.74	92.26

4.2 Human Posture Detection and Modeling Experiment

The relative positions and occlusion patterns of targets vary greatly between different viewpoints, and there are more ID matching errors and omissions in the detection algorithms. In the context of dense pedestrians under binocular vision, we addresses the issues of pedestrian pose point detection and association matching, jitter in continuous modeling, and local pose distortion. Therefore, a few more critical images in the video stream are extracted to demonstrate the operation effect of our method.

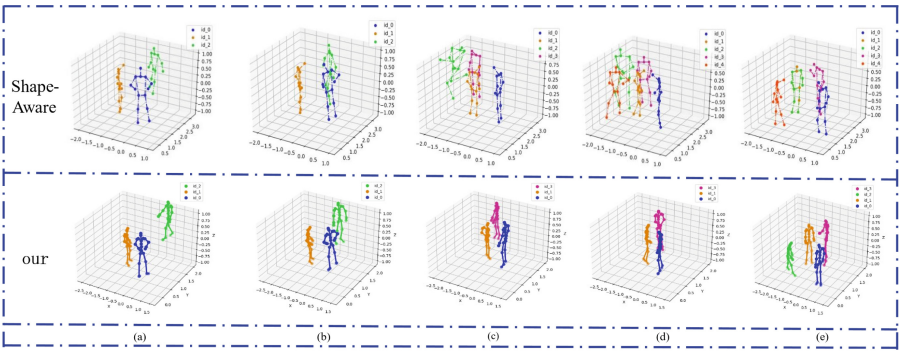


Fig. 6. 3D pose detection result in key frame

As shown in Fig. 6, comparing the algorithm we proposed with Shape-Aware, the framework is able to stably and accurately detect human 3D pose points when dealing with complex human interactions, and the algorithm still achieves better results even when there is an occlusion or overlapping of pedestrians, but Shape-Aware may produce different types of errors, such as (c) in which Shape-Aware gets a distorted pose detection that is not realistic when dealing with the person in the border, and (d) and (e) both of which appear to match the detected pose points to the wrong pedestrian target.

As shown in Fig. 7, the results of comparative and ablation experiments in modeling are presented. In subfigure (a), the 3D human body models constructed using the Shape-Aware and SPEC methods exhibited localized postural distortions, such as upper body deformation, leg shape distortion, and abnormal twisting of the feet. Although the

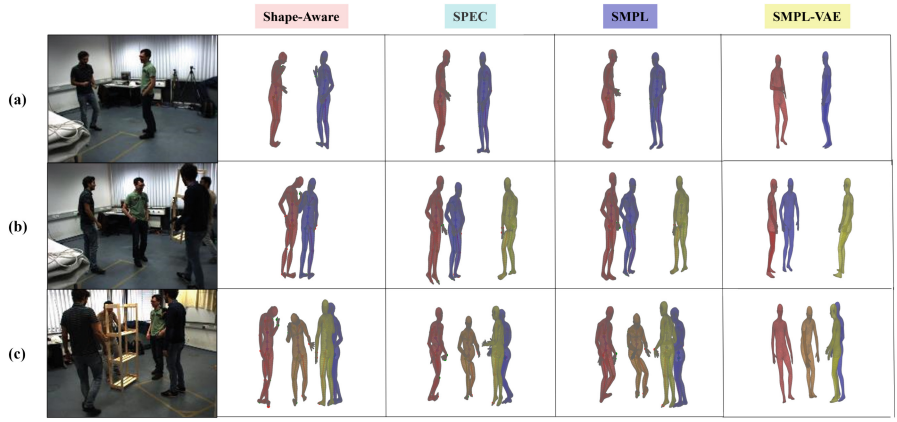


Fig. 7. Human body 3D modeling results

SMPL method improved the modeling results to some extent, significant corrections were achieved when incorporating the VAE module into the SMPL-VAE method. In subfigure (b), the Shape-Aware method exhibited omissions, and the SPEC and SMPL methods resulted in unnatural twists in the foot and head regions, respectively. These issues were effectively optimized in the modeling process using SMPL-VAE. In subfigure (c), problems in the detection of human posture points using Shape-Aware led to erroneous modeling results; SPEC and SMPL exhibited various degrees of modeling distortions (such as the orange model's abdomen and the blue pedestrian's upper body) that did not align with the actual human form, all of which were improved in the SMPL-VAE approach.

5 Conclusion

The 3D pose point detection method for the human body interior based on the skinned multiplayer linear model (SMPL) achieves excellent performance, with a value of 101.74 in the average per-joint position error metric and 92.26 in the percentage of correct keypoint metric, and also matches accurate pose points for different targets. For the noise generated by the clustering algorithm that maps the 2D pose points inside the human body from multiple viewpoints to the 3D pose points in the real scene, we remove the noise by means of Kalman filtering, which effectively solves the problem of jitter in the continuous modeling of pedestrians. The algorithm is based on the consideration of global optimization modeling results on the basis of focusing on the optimization of the local posture, through the automatic variational enconder (VAE) to the gradient descent regression network to correct the construction of the end-to-end human body 3D modeling network (SMPL-VAE), in order to maintain the overall proportion at the same time, more in line with the human body movement structure of the local modeling.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (52202417); China Postdoctoral Science Foundation (2022TQ0155,2022M721605); Open

Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (VRLAB2023A02); Young Elite Scientists Sponsorship Program by CAST (2023QNRC001); Young Elite Scientists Sponsorship Program by JSTJ (JSTJ-2023-XH032). Authors thank reviews for their valuable comments.

References

1. Cao, Z., Hidalgo, M.G., Simon, T.: OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 172–186 (2021)
2. Zhang, W.Q., Fang, J., Wang, X.G.: EfficientPose: efficient human pose estimation with neural architecture search. Comput. Vis. Media **7**, 335–347 (2021)
3. Sun, K., Xiao, B., Liu, D.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, NJ, pp. 5693–5703. IEEE (2019)
4. Toshev, A., Szegedy, C.: DeepPose: human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, NJ, pp. 1653–1660. IEEE (2014)
5. Sun, Y., Ye, Y., Liu, W.: Human mesh recovery from monocular images via a skeleton-disentangled representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, NJ, pp. 5349–5358. IEEE (2019)
6. Pavlakos, G., Choutas, V., Ghorbani, N.: Expressive body capture: 3D hands, face, and body from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, NJ, pp. 10975–10985 (2019)
7. Liu, S.C., Saito, S., Chen, W.K.: Learning to infer implicit surfaces without 3D supervision. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
8. Fang, H.S., Xie, S., Tai, Y.W.: RMPE: regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision, NJ, pp. 2334–2343. IEEE (2017)
9. Li, W.B., Wang, Z., Yin, B.Y., et al.: Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148 (2019)
10. Dong, Z.J., Song, J., Chen, X.: Shape-aware multi-person pose estimation from multi-view images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, NJ, pp. 11158–11168. IEEE (2021)
11. Loper, M., Mahmood, N., Romero, J.: SMPL: a skinned multi-person linear model. Seminal Graph. Pap.: Push. Bound. **2**, 851–866 (2023)
12. Li, Z., Oskarsson, M., Heyden, A.: 3D human pose and shape estimation through collaborative learning and multi-view model-fitting. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, NJ, pp. 1888–1897. IEEE (2019)
13. Diederik, P.K., Max, W.: Auto-encoding variational bayes. Comput. Sci. (2013)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. Commun. ACM **60**(6), 84–90 (2017)
15. Ishwarya, K., Nithya, A.A.: Squirrel search optimization with deep convolutional neural network for human pose estimation. Comput. Mater. Continua **74**(3) (2023)
16. Tekin, B., Katircioglu, I., Salzmann, M.: Structured prediction of 3D human pose with deep neural networks. arXiv preprint arXiv:1605.05180 (2016)
17. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, NJ, pp. 10133–10142. IEEE (2019)

18. Tu, H., Wang, C., Zeng, W.: VoxelPose: towards multi-camera 3D human pose estimation in wild environment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, JM. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 197–212. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_12
19. Bogo, F., Kanazawa, A., Lassner, C.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part V. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
20. Kanazawa, A., Black, M.J., Jacobs, D.W.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, NJ, pp. 7122–7131. IEEE (2018)
21. Kwon, O.-H., Tanke, J., Gall, J.: Recursive Bayesian filtering for multiple human pose tracking from multiple cameras. In: Ishikawa, H., Liu, C.-L., Pajdla, T., Shi, J. (eds.) ACCV 2020. LNCS (LNAI and LNB), vol. 12623, pp. 438–453. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69532-3_27
22. Tulsiani, S, Efros, A.A., Malik, J.: Multi-view consistency as supervisory signal for learning shape and pose prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition, NJ, pp. 2897–2905. IEEE (2018)
23. Ke, S.R., Zhu, L.J., Hwang, J.N.: Real-time 3D human pose estimation from monocular view with applications to event detection and video gaming. In: 2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, NJ, pp. 489–496. IEEE (2010)
24. Shi, J.R., Wang, D., Shang, F.H.: Research progress of stochastic gradient descent algorithm. Acta Autom. Sinica **47**(9), 2103–2119 (2019)
25. Wei, L., Zheng, C., Hu, Y.: Oriented object detection in aerial images based on the scaled smooth L1 loss function. Remote Sens. **15**(5), 1350 (2023)
26. Ramakrishna, V., Munoz, D., Hebert, M.: Pose machines: articulated pose estimation via inference machines. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 33–47. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-319-10605-2_3
27. Belagiannis V, Amin S, Andriluka M.: 3D pictorial structures for multiple human pose estimation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, vol. 1, pp. 1669–1676. IEEE Computer Society (2014)
28. Tian, W, Gao, Z, Tan, D.: Single-view multi-human pose estimation by attentive cross-dimension matching. Front. Neurosci. **17** (2023)
29. Rani, C.J., Devarakonda, N., Kumari, K.W.S.N., Malavath, P.: A monadic and effective frame work for single human pose estimation of 2D images and videos. In: Chen, J.I.Z., Tavares, J.M.R.S., Iliyasu, A.M., Du, K.L. (eds.) ICIPCN 2021. LNNS, vol. 300, pp. 254–268. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-84760-9_23