Available Online: 1 April 2026 DOI: 10.54254/2977-3903/2026.32472

Multi-view 3D human pose estimation based on geometric constraints

Yang Lin*, Lijing Tong, Zheng Han

North China University of Technology, Beijing, China

*Corresponding Author. Email: 3478646436@qq.com

Abstract. In the field of computer vision, the existing real-time multi-view 3D human pose estimation method Faster VoxelPose (FVP) suffers from redundant detections and fixed plane-fusion weights in densely occluded scenarios. To address these issues, this study proposes an improved method based on Geometric Constraints, termed FVP-GC. During the human detection stage, a Redundancy Suppression Mechanism (RSM) is introduced. This mechanism applies geometric constraint filtering to predicted bounding boxes during training using dual thresholds of distance and overlap ratio, thereby reducing redundant detections caused by projection overlaps. During the keypoint localization stage, a Geometric Consistency Adaptive Weighting (GCAW) strategy is proposed. By evaluating the geometric consistency among orthogonal planes, the method dynamically adjusts fusion weights, improving the fusion quality of multi-plane predictions. Experimental results on the Shelf and CMU Panoptic datasets demonstrate that, while maintaining real-time inference speed, the proposed method improves pose estimation accuracy compared with Faster VoxelPose, achieving average Percentage of Correct Parts in 3D (PCP3D) values of 97.7% and 97.1%, and reducing Mean Per Joint Position Error (MPJPE) to 17.45 mm.

Keywords: 3D human pose estimation, multi-view, Faster VoxelPose, geometric constraints, geometric consistency

1. Introduction

With the rapid development of artificial intelligence and computing power, human pose estimation [1] has been widely applied in fields such as human–computer interaction [2], motion capture [3], virtual avatars [4], rehabilitation training [5], sports performance analysis [6], and construction safety. However, real-world environments often involve occlusion, dense crowd interactions, and dynamic background interference, which significantly constrain the performance of existing methods. Therefore, achieving robust and real-time multi-view 3D human pose estimation in multi-person scenarios remains a critical challenge in the field of computer vision.

At present, multi-view 3D human pose estimation in multi-person scenes mainly faces two major challenges: first, the accurate association of joints and individuals across different views; and second, the handling of mutual occlusions in dense crowds. To address the cross-view association problem, researchers have proposed strategies including re-identification features [7], dynamic matching [8, 10], 4D graph cuts

[11], plane-sweeping stereo [12], and ground projection based on threshold clustering [13]. However, these methods are susceptible to noise from two-dimensional pose estimation in crowded environments. To tackle occlusion, Belagiannis et al. compressed the state space of the 3D PSM through multi-view triangulation and introduced probabilistic models for occlusion and cross-view ambiguity, thereby achieving preliminary multi-person 3D pose estimation. Chen et al. [14] incorporated temporal information into pose estimation to improve joint localization accuracy. Wang et al. [15] proposed a direct regression model integrating projection and Transformer architectures. Srivastav et al. [16] employed adaptive supervised attention to alleviate pseudo-label noise. Nevertheless, these approaches rely heavily on camera spatial configurations, and their generalization ability in complex environments remains limited.

In recent years, voxel-based 3D human pose estimation methods [17, 22] have gradually become a major research focus. These methods back-project multi-view two-dimensional features into a unified three-dimensional voxel space and perform inference directly in the 3D domain, fundamentally avoiding the ambiguity of cross-view matching and demonstrating clear advantages in occlusion scenarios. Tu et al. [17] first proposed VoxelPose, which achieves robust multi-person pose estimation through three-dimensional voxel feature aggregation and a two-stage network. Deng et al. [18] introduced domain adaptation and transferable parameter learning to enhance the model's robustness to domain shifts. Zhang et al. [19] proposed VoxelTrack, which jointly optimizes three-dimensional joint coordinates and human appearance embeddings, enabling simultaneous pose estimation and identity-consistent tracking. However, these high-precision approaches rely on computationally intensive three-dimensional convolutions, making it difficult to satisfy real-time requirements. To overcome this speed bottleneck, Chen et al. [20] reduced voxel computation costs using sparse self-attention, while Ye et al. [21] proposed Faster VoxelPose. The latter compresses three-dimensional voxel features into three two-dimensional planes via orthogonal projection and replaces three-dimensional convolutions with efficient two-dimensional convolutions, significantly improving inference speed while maintaining accuracy. Zhuang et al. [22] further introduced a depth projection attenuation mechanism to mitigate the depth ambiguity caused by orthogonal projection.

Despite these advances, Faster VoxelPose still exhibits two aspects that can be further optimized in densely occluded scenarios. First, orthogonal projection causes individual bounding boxes to overlap easily in the two-dimensional planes, leading to redundant detections that interfere with subsequent human localization. Second, the prediction weights of orthogonal planes are relatively fixed and cannot dynamically adjust the fusion strategy according to the actual reliability of each plane, leaving room for improvement in the robustness of joint localization. To address these issues, this paper proposes an improved method based on geometric constraints, termed Faster VoxelPose with Geometric Constraints (FVP-GC). During the human detection stage, a Redundancy Suppression Mechanism (RSM) is introduced, which applies geometric constraint filtering to predicted bounding boxes during training using dual thresholds of distance and overlap ratio, thereby reducing redundant detections caused by projection overlap. During the keypoint localization stage, a Geometric Consistency Adaptive Weighting (GCAW) strategy is incorporated. By dynamically adjusting fusion weights according to the geometric consistency among orthogonal planes, the method improves the fusion quality of multi-plane predictions in a lightweight manner. Experimental results on multiple public datasets demonstrate that the proposed method improves pose estimation accuracy compared with Faster VoxelPose while maintaining real-time inference speed, confirming its effectiveness and practical potential in complex scenarios.

2. FVP-GC network

To better illustrate the proposed FVP-GC network, its overall framework is shown in Figure 1.

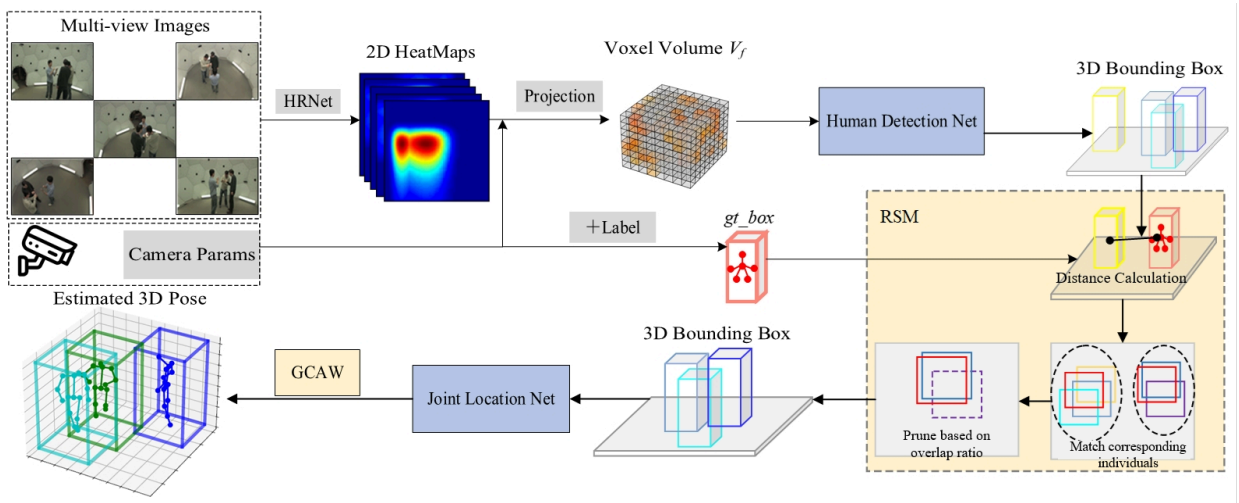

Figure 1. Architecture of the FVP-GC network

During the voxel space construction stage, two-dimensional pose heatmaps are first estimated from multi-view images using the pre-trained HRNet [23] These heatmaps are then back-projected into a three-dimensional voxel space $V_f \in R^{K\times L\times W\times H}$ (where L, W, H denote the voxel volume dimensions, and J represents the number of joints). Subsequently, a human detection network is employed to generate initial three-dimensional human bounding boxes. To further eliminate redundancy and false detections caused by projection overlap in dense scenes, this study introduces a Redundancy Suppression Mechanism (RSM) during the training phase. RSM applies geometric constraint filtering to candidate bounding boxes using dual thresholds based on distance and overlap ratio. This process provides cleaner inputs for subsequent localization while reducing computational overhead. Next, the joint localization network processes the voxel features within each three-dimensional bounding box through a lightweight convolutional network to generate two-dimensional joint heatmaps on the corresponding orthogonal planes. Finally, to address the unreliability of single-plane predictions caused by occlusion, a Geometric Consistency Adaptive Weighting (GCAW) strategy is proposed. By exploiting geometric consistency among orthogonal planes, the method dynamically evaluates the prediction confidence of each plane and performs adaptive fusion to obtain the final three-dimensional joint coordinates. While maintaining real-time inference speed, FVP-GC effectively alleviates the problems of detection redundancy and fusion bias in densely occluded scenarios, thereby improving the accuracy and robustness of pose estimation.

2.1. Redundancy suppression mechanism

Although the Faster VoxelPose network improves human localization quality through feature purification, dense occlusion scenarios still present challenges. Due to orthogonal projection, individual bounding boxes tend to overlap heavily on the two-dimensional planes. Combined with residual feature noise, two types of undesirable outputs may still occur during the detection stage: False detection boxes located far from the actual human body. Redundant detection boxes that correspond to the same real human but overlap heavily with one another. Figure 2 shows a detection example from a frame in the Shelf dataset. In Figure 2(c), the elliptical region in the upper-left corner shows a false skeleton generated by background interference, while

the elliptical region in the lower-left corner illustrates redundant skeletons produced by dense crowds and projection overlap. These extra detection boxes compete with real human instances for computational resources, interfere with the accuracy of subsequent joint localization, and may even disrupt the spatial topological structure and identity consistency of pose estimation.

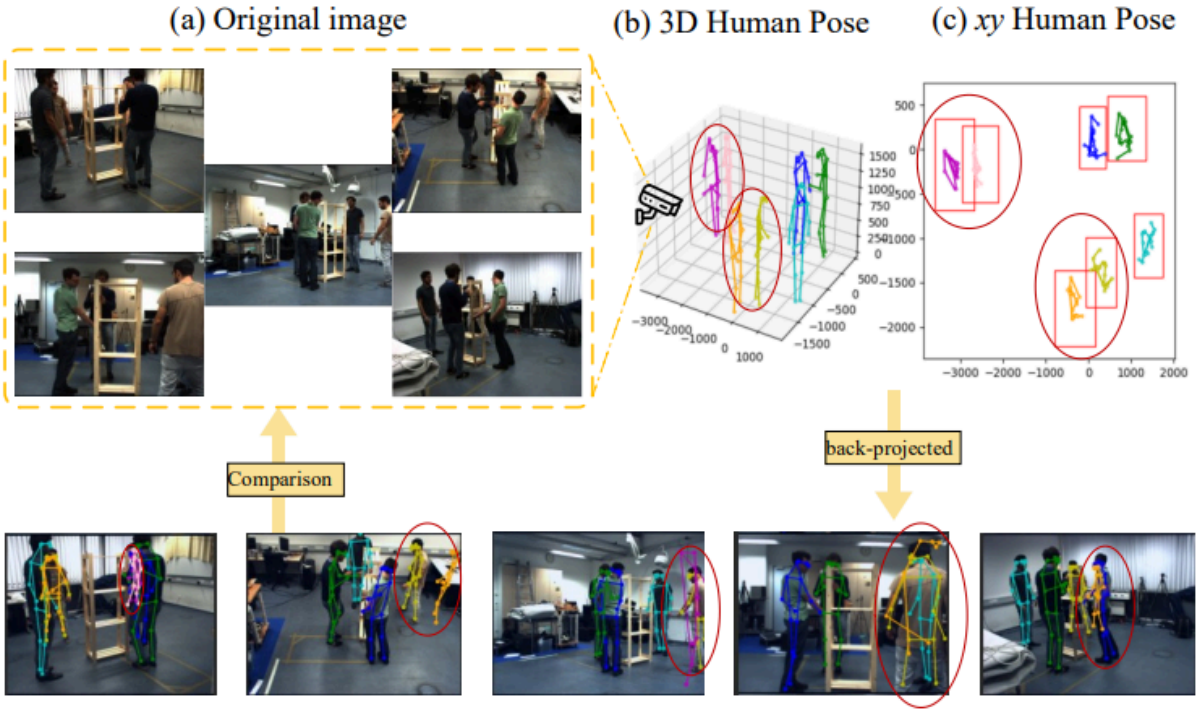

Figure 2. Examples of redundant and false detection boxes in human pose estimation

To address this problem, this paper proposes a Redundancy Suppression Mechanism (RSM). RSM is a deterministic post-processing algorithm based on geometric constraints that is applied during the training phase. Its core idea is to use ground-truth annotations to perform geometric constraint screening and matching for candidate three-dimensional bounding boxes, while assigning maximum loss to clearly incorrect or redundant predictions. This process guides the detection network to produce cleaner and more accurate outputs.

The RSM procedure can be summarized in three key steps:

(1) False box filtering: The three-dimensional Euclidean distance between each candidate box and all ground-truth boxes is computed, and the nearest ground-truth box is assigned. If the minimum distance exceeds the preset threshold of 500 mm, the candidate box is regarded as a false detection and its matching relationship is marked as invalid.

(2) Size correction: For successfully matched candidate boxes, if the predicted size is smaller than that of the corresponding ground-truth box, the size is forcibly adjusted to match the ground-truth box, preventing missed detections caused by overly small boxes.

(3) Overlap suppression: If multiple candidate boxes match the same ground-truth box and their pairwise 3D Intersection over Union (IoU) exceeds the threshold of 25%, only the candidate box with the highest confidence is retained, while the others are marked as redundant and invalidated.

The above thresholds are determined through sensitivity analysis experiments, achieving an optimal balance between noise filtering and valid detection preservation. The structure of the RSM module is illustrated in Figure 3. During forward propagation, RSM does not participate in computation and does not

modify the network structure. Instead, prior to backpropagation, the matching matrix produced by RSM is used to mask the loss function so that the loss is computed only from purified high-quality predictions. Since RSM involves only distance calculation, threshold comparison, and indexing operations, it introduces no additional computational overhead during inference. By reducing redundant and false detections at the source, RSM effectively purifies the input to the subsequent localization module and significantly improves the accuracy and efficiency of human detection in complex scenes.

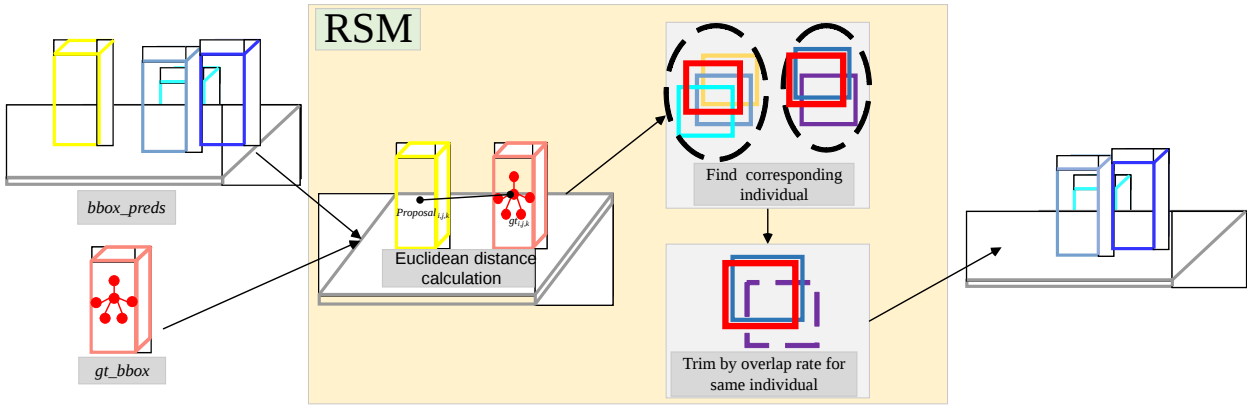

Figure 3. Structure of the RSM module

2.2. Geometric consistency adaptive weighting

During the joint localization stage, Faster VoxelPose orthogonally projects the voxel features within each three-dimensional bounding box onto the three planes xy, xz, and yz. A lightweight convolutional network is then used to generate two-dimensional joint heatmaps on each plane. However, due to factors such as the degree of occlusion, projection angle, and feature quality, the prediction reliability of the three planes may vary. The fixed-weight fusion strategy adopted by Faster VoxelPose cannot adapt to such variations, leaving room for further improvement in the robustness of joint localization. To address this limitation, this study introduces a Geometric Consistency Adaptive Weighting (GCAW) strategy. The working mechanism of this strategy is illustrated in Figure 4.

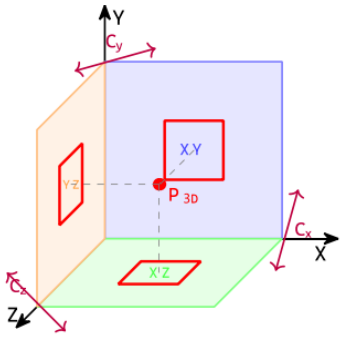

Figure 4. Mechanism of the GCAW strategy

The basic idea is to use the inherent geometric relationships among the three orthogonal planes as a reference for prediction confidence. Ideally, the xy and xz planes share the same X coordinate; similarly, the xy and yz planes share the same Y coordinate, while the xz and yz planes share the same Z coordinate. When occlusion or feature degradation causes deviations in the prediction of a single plane, the

consistency of the predicted coordinate component shared with the corresponding plane will also decrease. GCAW quantifies this geometric consistency to dynamically evaluate the reliability of each plane, thereby achieving adaptive weighted fusion in a lightweight manner.

Specifically, for the i -th human instance, let the predicted joint coordinates on the three planes be $P_{xy} \in R^{J\times 2}$, $P_{xy} \in R^{J\times 2}$, $P_{yz} \in R^{J\times 2}$, respectively. The geometric consistency score for each coordinate dimension is computed as shown in the following Equation (1).

$$C_x = e^{\frac{-|X_{xy}-X_{xz}|}{\tau}}, C_y = e^{\frac{-|X_{xy}-X_{yz}|}{\tau}}, C_z = e^{\frac{-|X_{xz}-X_{yz}|}{\tau}} \tag{1}$$

τ= 1 is used to regulate the sensitivity of the consistency score. This function maps the prediction deviation into the interval (0,1]. A smaller deviation results in a higher consistency score, indicating that the predictions of the two planes corresponding to that coordinate dimension are closer to each other.

Based on these geometric consistency scores, GCAW dynamically computes fusion weights for each coordinate dimension. Taking the X coordinate as an example, which is jointly determined by the xy and xz planes, the weight calculation is defined as Equation (2)~(3).

$$\widetilde{w}_{xy}^{x} = w_{xy} \cdot (1 + \alpha \cdot C_x), \widetilde{w}_{xz}^{x} = w_{xz} \cdot (1 + \alpha \cdot C_z) \tag{2}$$

$$w_{xy}^{x} = \frac{\widetilde{w}_{xy}^{x}}{\widetilde{w}_{xy}^{x}+\widetilde{w}_{xz}^{x}+\varepsilon}, w_{xz}^{x} = \frac{\widetilde{w}_{xz}^{x}}{\widetilde{w}_{xy}^{x}+\widetilde{w}_{xz}^{x}+\varepsilon} \tag{3}$$

where w_{xy} , w_{xz} denote the base weights predicted by the network, α represents the geometric consistency adjustment coefficient, and $\varepsilon = 1 \times 10^{-6}$ is a numerical stability term. The weight calculations for the Y and Z coordinates follow the same principle. The final three-dimensional pose $P = [X, Y, Z] \in R^{J\times 3}$ is obtained through adaptive weighted fusion, as expressed in the following Equation (4)~(6).

$$X = w_{xy}^{x} \cdot X_{xy} + w_{xz}^{x} \cdot X_{xz} \tag{4}$$

$$Y = w_{xy}^{y} \cdot Y_{xy} + w_{yz}^{y} \cdot Y_{yz} \tag{5}$$

$$Z = w_{xz}^{z} \cdot X_{xz} + w_{yz}^{z} \cdot X_{yz} \tag{6}$$

The GCAW strategy involves only element-wise operations and matrix multiplications, introducing no additional parameters. Its computational complexity grows linearly with the number of joints and has minimal impact on the overall inference speed. By explicitly modeling the geometric consistency constraints among the three planes, GCAW dynamically adjusts fusion weights according to the reliability of each plane's prediction. This effectively mitigates the influence of occlusion or feature degradation in individual planes and yields more stable and accurate joint localization results in complex scenarios.

2.3. Loss function

The loss function of the proposed FVP-GCnetwork, denoted as Loss, is composed of three components: the 1D bounding box loss L_{1d} , the 2D bounding box loss L_{2d} , and the joint loss L_{joints} . The overall loss is obtained by summing these three terms.

L_{1d} is used to measure the error between the predicted height of a bounding box and the ground-truth height. Its computation is defined as Equation (7)~(8).

$$\widehat{H}_{n,d}^{(z)} = max_{1\le n\le N} exp\{-\frac{(d-\widehat{d}_n)^2}{2\sigma^2}\} \tag{7}$$

$$L_{1d} = \frac{1}{N}\left(\sum_{n=1}^{N}\sum_{d=1}^{D}\|H_{n,d}^{(z)} - \widehat{H}_{n,d}^{(z)}\|_2\right)_{Mask} \tag{8}$$

where N denotes the preset maximum number of persons in the scene, and $\widehat{d}_n$ represents the ground-truth position of the n-th individual. The term $(\cdot)_{Mask}$ denotes the 1D bounding box loss computed using a masking mechanism based on the matching matrix generated by RSM. D represents the maximum height of the voxel space. The indicator function $H_{n,d}^{(z)}$ indicates whether the n-th person exists at height position d, while $\widehat{H}_{n,d}^{(z)}$ denotes the confidence score that the nnn-th person is located at height d.

L_{2d} measures the localization error of the human body center as well as the discrepancy between the predicted bounding box size on the xy plane and the corresponding ground-truth size. The computation is defined as Equation (9).

$$L_{2d} = \begin{cases} \sum_{i=1}^{L}\sum_{j=1}^{W}(\|H_{i,j}^{(xy)} - \widehat{H}_{i,j}^{(xy)}\|_2 + \frac{\lambda_{box}}{N}\|S_{i,j} - \widehat{S}_{i,j}\|_1), Mask_{i,j} = 1 \\ \sum_{i=1}^{L}\sum_{j=1}^{W}(\|H_{i,j}^{(xy)}\|_2 + \frac{\lambda_{box}}{N}\|S_{i,j}\|_1), Mask_{i,j} = 0 \end{cases} \tag{9}$$

where L, W denote the maximum length and width of the voxel space, respectively. The terms $H_{i,j}^{(xy)}$ and $\widehat{H}_{i,j}^{(xy)}$ have meanings similar to $H_{n,d}^{(z)}$ and $\widehat{H}_{n,d}^{(z)}$ in the previous equation, except that they represent the probability of the n -th person appearing at position (i, j) on the xy plane. The parameter λ_{box} serves as a weighting coefficient. $S_{i,j}$ and $\widehat{S}_{i,j}$ represent the sizes of the ground-truth bounding box and the predicted bounding box on the xy plane, respectively. The indicator variable $Mask_{i,j} = 1$ means the position is valid; otherwise it is not.

L_{joints} measures both the discrepancy between the predicted joint positions on the three orthogonal planes and the corresponding ground-truth annotations, as well as the error between the predicted 3D pose and the ground-truth pose. The calculation is given as Equation (10).

$$L_{joints} = (\sum_{n=1}^{N}(\sum_{t}\|J_n^{(t)} - \widehat{J}_n^{(t)}\|_1 + \lambda_{fused}\|P_n - \widehat{P}_n\|_1))_{Mask} \tag{10}$$

where $(\cdot)_{Mask}$ is computed using the proposal obtained through the masking mechanism. N denotes the total number of human body joints, and $t \in \{xy, xz, yz\}$ represents the three orthogonal planes. $J_n^{(t)}$ and $\widehat{J}_n^{(t)}$ denote the ground-truth and predicted positions of the n-th joint on plane t, respectively. λ_{fused} is a weighting coefficient. P_n and $\widehat{P}_n$ denote the ground-truth and predicted positions of the n-th joint in the fused three-dimensional pose.

3. Experimental results and comparative analysis

3.1. Experimental setup and datasets

All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 2080 Ti GPU, using PyTorch 2.0 as the deep learning framework. The network was trained using the Adam optimizer with a batch size of 8.

Two publicly available multi-view datasets, Shelf and CMU Panoptic, were used for evaluation. The Shelf dataset consists of a dense four-person assembly/disassembly scenario captured by five calibrated cameras, where severe occlusions frequently occur. The CMU Panoptic dataset [24] is a large-scale multi-view dataset that includes 480 VGA cameras and 31 high-definition cameras. The training–testing splits of all datasets follow the configuration described in Reference [25].

3.2. Evaluation metrics

To comprehensively evaluate the performance of the proposed method, different evaluation metrics were adopted according to the characteristics of each dataset, and additional efficiency metrics were introduced to assess the practicality and real-time capability of the model.

In terms of accuracy evaluation, the Percentage of Correct Parts in 3D (PCP_{3D}) was used for the Shelf and Campus datasets. This metric measures whether the predicted error of skeletal endpoints is less than 50% of the ground-truth limb length, thereby assessing the overall accuracy of human pose estimation.For the CMU Panoptic dataset, two metrics were employed: Average Precision (AP) and Mean Per Joint Position Error (MPJPE). AP reflects the accuracy of three-dimensional human detection by integrating precision across different recall levels. MPJPE calculates the average Euclidean distance between predicted joint positions and the corresponding ground-truth locations, directly measuring the accuracy of joint localization.

In terms of efficiency evaluation, several indicators were introduced to verify the model's real-time processing capability and computational complexity. Frames Per Second (FPS) measures inference speed; Multiply–Accumulate Operations (MACs) and the number of parameters (Parameters) evaluate the computational and storage costs of the model; and single-frame inference time (Time) reflects the actual processing time required for each input frame.

3.3. Experimental results and analysis

3.3.1. Ablation study

To verify the effectiveness of the proposed RSM and GCAW modules, Faster VoxelPose was used as the baseline model. Each module was incrementally integrated into the baseline for ablation experiments. The results are presented in Table 1.

Table 1. Ablation study results

	Model	AVG PCP_{3D} (%)	MPJPE (mm)
		Shelf	Panoptic
A	Baseline	97.40	18.26
B	A+RSM	97.68	17.57
C	A+GCAW	97.63	17.72
D	B+GCAW	97.73	17.45

As shown in Table 1, the baseline Faster VoxelPose (Model A) achieves a PCP_{3D} of 97.40% on the Shelf dataset and an MPJPE of 18.26 mm on the CMU Panoptic dataset. After introducing the RSM module alone (Model B), the two metrics improve to 97.68% and 17.57 mm, respectively, indicating that the redundancy suppression mechanism effectively refines detection results. When only the GCAW module is introduced (Model C), PCP_{3D} increases to 97.63% and MPJPE decreases to 17.72 mm, demonstrating the positive effect of geometric-consistency-based weighting on joint localization. When both RSM and GCAW are integrated (Model D, i.e., FVP-GC), the best performance is achieved. The PCP_{3D} reaches 97.73%, and MPJPE decreases to 17.45 mm, representing improvements of 0.33% and 0.81 mm, respectively, compared with the baseline. These results confirm the complementary and synergistic effectiveness of the two modules.

3.3.2. Comparative experiments

To further validate the effectiveness of FVP-GC, comparisons were conducted with several representative methods proposed in recent years on the Shelf and CMU Panoptic datasets.

As shown in Table 2, on the Shelf dataset, which represents an indoor scenario with dense occlusions, FVP-GC achieves an average PCP3D of 97.7%, which is comparable to the current state-of-the-art methods. Notably, for Actor 2, who experiences the most severe occlusion, FVP-GC achieves an accuracy of 96.5%, significantly outperforming the baseline Faster VoxelPose (95.4%) as well as most of the comparison methods. This result demonstrates the effectiveness of RSM in suppressing redundant detections and GCAW in enabling adaptive fusion. Moreover, FVP-GC exhibits relatively balanced performance across the three actors (Actor 1: 98.9%, Actor 2: 96.5%, Actor 3: 97.7%), without any obvious performance weaknesses. This indicates strong robustness and generalization capability in complex occlusion scenarios.

Table 2. Comparative results on the shelf dataset

Network Name	Shelf-PCP_{3D} (%)			
	Act.1	Act.2	Act.3	Avg
Dong et al. [7]	98.8	94.1	97.8	96.9
Huang et al. [26]	98.8	96.2	97.2	97.4
VoxelPose [17]	99.3	94.1	97.6	97.0
VoxelTrack [19]	98.6	94.9	97.7	97.1
MvP [15]	99.3	95.1	97.8	97.4
Baseline [21]	99.1	95.4	97.7	97.4
Chen et al. [14]	98.6	95.8	97.9	97.4
VTP [20]	99.3	95.1	97.4	97.3
Deng et al. [18]	96.5	94.1	97.7	96.1
SelfPose3d [16]	97.2	90.3	97.9	95.1
FVP-GC	98.9	96.5	97.7	97.7

The experimental results on the CMU Panoptic dataset are presented in Table 3. FVP-GC achieves 88.82% on the AP_{25} metric, representing an improvement of 3.60% over the baseline Faster VoxelPose (85.22%), and outperforming mainstream methods such as VoxelPose (83.59%) and VTP (83.79%). This demonstrates that the proposed modules effectively enhance detection accuracy in complex scenarios. In terms of joint localization accuracy, MPJPE decreases by 0.81 mm compared with the baseline, reaching 17.45 mm, which is lower than that of most comparison methods. This result reflects the advantage of the proposed approach in precise joint localization. Regarding efficiency, the computational complexity of FVP-GC increases slightly compared with the baseline. The additional computational overhead mainly originates from the calculation of geometric consistency scores and the dynamic adjustment of fusion weights in GCAW, which involve only element-wise operations and matrix multiplications. The additional MACs amount to approximately 0.14 G, accounting for about 0.4% of the total computation. Although the inference speed decreases from 31.1 FPS in the baseline to 29.3 FPS (a reduction of approximately 5.8%), it still exceeds the threshold for real-time processing (> 25 FPS). Combined with the improvement in MPJPE, these results indicate that GCAW achieves significant accuracy gains with minimal computational overhead, striking a favorable balance between accuracy and efficiency and demonstrating its potential for practical applications.

Table 3. Comparative results on the CMU panoptic dataset

Network Name	AP_{25}	AP_{50}	AP_{100}	AP_{150}	MPJPE	MACs	Params	Time	FPS
VoxelPose [17]	83.59	98.33	99.76	99.91	17.68	-	-	316.0	3.2
VoxelTrack [19]	79.34	96.83	99.58	-	18.49	-	-	-	-
Faster VoxelPose [21]	85.22	98.08	99.32	99.48	18.26	34.91	2.64	32.2	31.1
VTP [20]	83.79	97.14	98.15	98.40	17.62	-	-	-	-
Zhu et al. [27]	61.28	95.10	99.39	99.87	18.88	-	-	-	-
FVP-GC	88.82	98.34	99.42	99.73	17.45	35.05	2.67	34.1	29.3

3.3.3. Visualization

To visually demonstrate the performance improvements achieved by the proposed method, three-dimensional pose estimation results on the Shelf and CMU Panoptic datasets are presented for comparison, as illustrated in Figure 5 and 6, respectively.

As shown in Figure 5, in the densely occluded scenes of the Shelf dataset, the baseline method Faster VoxelPose exhibits joint localization deviations and redundant detections in certain occluded regions. In contrast, FVP-GC produces more accurate pose estimations. The joint positions in occluded areas better conform to the human body structure, and redundant predictions are effectively reduced.

In the complex scenarios of the Panoptic dataset, as illustrated in Figure 6, FVP-GC also demonstrates superior localization accuracy. Compared with the baseline method, where certain joints show noticeable positional deviations, the predictions of FVP-GC align more closely with the true human body structure. These results further verify the effectiveness and generalization capability of the proposed method in complex environments.

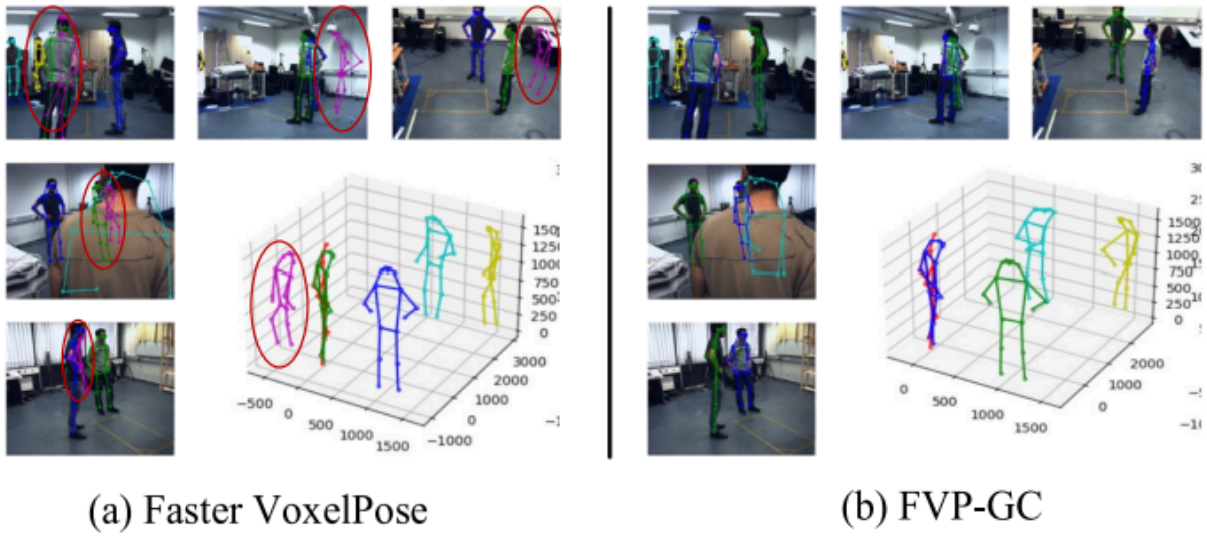

Figure 5. Visualization comparison of 3D pose estimation on the Shelf dataset

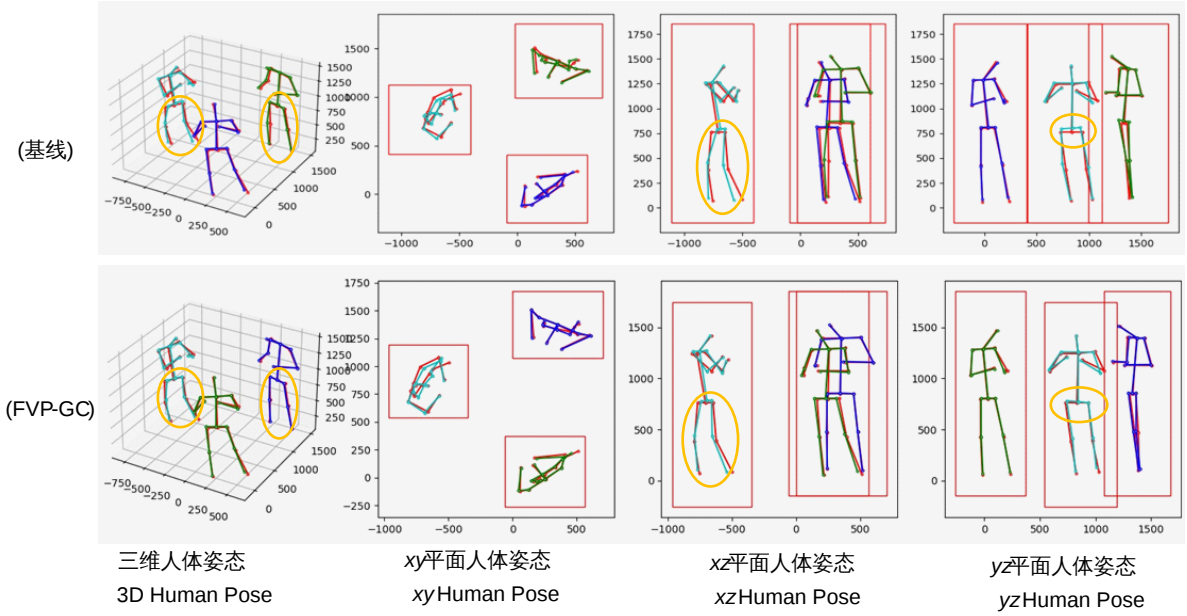

Figure 6. Visualization comparison of 3D pose estimation on the Panoptic dataset

4. Conclusion

To address the issues of redundant detections and fixed fusion weights in Faster VoxelPose under densely occluded scenarios, this paper proposes an improved method based on geometric constraints, termed FVP-GC. The proposed approach introduces a Redundancy Suppression Mechanism (RSM) to filter out false and redundant detection boxes, and employs a Geometric Consistency Adaptive Weighting (GCAW) strategy to dynamically fuse predictions from three orthogonal planes. These mechanisms effectively enhance the accuracy of joint localization. Experimental results on the Shelf and CMU Panoptic datasets demonstrate that FVP-GC improves pose estimation accuracy compared with the baseline method while maintaining real-time inference speed, thereby confirming its effectiveness in complex scenarios. Future work may further explore the following two directions. First, the temporal stability of the RSM and GCAW strategies in dynamic video sequences could be investigated. For example, lightweight temporal modules could be introduced to perform temporal smoothing of geometric consistency scores, thereby mitigating the influence of occasional single-frame prediction deviations on overall pose estimation. Second, the adaptability of the proposed method in scenarios involving severe non-rigid deformations, such as sports competitions or dance movements, warrants further study. In such cases, rapid motion may introduce projection distortions and occlusion effects that disrupt geometric consistency among orthogonal planes. Accordingly, more robust geometric constraint mechanisms could be designed to improve the model's generalization capability in more challenging environments.

Funding project

Open Project of the Chinese Academy of Cultural Heritage (2025KF11)

References

[1] Nogueira, A. F. R., Oliveira, H. P., & Teixeira, L. F. (2025). Markerless multi-view 3D human pose estimation: A survey. *Image and Vision Computing, 149*, Article 105437. https: //doi.org/10.1016/j.imavis.2025.105437

[2] Chen, L., Ai, H., Chen, R., & Zhuang, Z. (2020). Cross-view tracking for multi-human 3D pose estimation at over 100 fps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 3279–3288). IEEE.

[3] Deng, X., Sheng, Y., Pei, H., & Guo, Y. (2024). Posture recognition method of duty personnel based on human posture key points and convolutional neural network. *Journal of Electronic Imaging, 33*(2), 023054. https: //doi.org/10.1117/1.JEI.33.2.023054

[4] Weng, C.-Y., Curless, B., & Kemelmacher-Shlizerman, I. (2019). Photo wake-up: 3D character animation from a single photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 5908–5917). IEEE.

[5] Du, H., Zhao, Y., Han, J., & Xu, D. (2016). Data fusion of human skeleton joint tracking using two Kinect sensors and extended set membership filter. *Acta Automatica Sinica, 42*(12), 1886–1898.

[6] Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., & Shao, L. (2021). Deep 3D human pose estimation: A review. *Computer Vision and Image Understanding, 210*, Article 103225. https: //doi.org/10.1016/j.cviu.2021.103225

[7] Dong, J., Jiang, W., Huang, Q., Bao, H., & Zhou, X. (2019). Fast and robust multi-person 3D pose estimation from multiple views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 7792–7801). IEEE.

[8] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., & Ilic, S. (2014). 3D pictorial structures for multiple human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 1669–1676). IEEE.

[9] Belagiannis, V., Amin, S., Andriluka, M., Schiele, B., Navab, N., & Ilic, S. (2016). 3D pictorial structures revisited: Multiple human pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(10), 1929–1942. https: //doi.org/10.1109/TPAMI.2015.2507125

[10] Belagiannis, V., Wang, X., Schiele, B., & Ilic, S. (2015). Multiple human pose estimation with temporally consistent 3D pictorial structures. In *Computer Vision – ECCV 2014 Workshops* (pp. 742–754). Springer.

[11] Zhang, Y., An, L., Yu, T., Li, X., Li, K., & Liu, Y. (2020). 4D association graph for realtime multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 1324–1333). IEEE.

[12] Lin, J., & Lee, G. H. (2021). Multi-view multi-person 3D pose estimation with plane sweep stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 11886–11895). IEEE.

[13] Chen, X., Wang, L., Tang, J., & Guo, Y. (2025). Real-time reconstruction of multiperson 3D human pose with perspective transformation matching based on multiview. *Journal of Electronic Imaging, 34*(2), 023023. https: //doi.org/10.1117/1.JEI.34.2.023023

[14] Chen, L., Liu, T., Gong, Z., Lu, M., & Liu, W. (2024). Movement function assessment based on human pose estimation from multi-view. *Computer Systems Science & Engineering, 48*(2), 321–339. https: //doi.org/10.32604/csse.2024.047234

[15] Zhang, J., Cai, Y., Yan, S., Feng, J., & et al. (2021). Direct multi-view multi-person 3D human pose estimation. In *Advances in Neural Information Processing Systems 34* (NeurIPS) (pp. 13153–13164). MIT Press.

[16] Srivastav, V., Chen, K., & Padoy, N. (2024). SelfPose3d: Self-supervised multi-person multi-view 3D pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 2502–2512). IEEE.

[17] Tu, H., Wang, C., & Zeng, W. (2020). VoxelPose: Towards multi-camera 3D human pose estimation in wild environment. In *Proceedings of the European Conference on Computer Vision* (ECCV) (pp. 197–212). Springer.

[18] Deng, J., Yao, H., & Shi, P. (2023). Enhanced 3D pose estimation in multi-person, multi-view scenarios through unsupervised domain adaptation with dropout discriminator. *Sensors, 23*(20), Article 8406. https: //doi.org/10.3390/s23208406

[19] Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2022). VoxelTrack: Multi-person 3D human pose estimation and tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(2), 2613–2626. https: //doi.org/10.1109/TPAMI.2021.3061489

[20] Chen, Y., Gu, R., Huang, Q., & Wang, F. (2023). VTP: Volumetric transformer for multi-view multi-person 3D pose estimation. *Applied Intelligence, 53*(22), 26568–26579. https: //doi.org/10.1007/s10489-023-04951-6

[21] Ye, H., Zhu, W., Wang, C., Wu, R., & Wang, X. (2022). Faster VoxelPose: Real-time 3D human pose estimation by orthographic projection. In *Proceedings of the European Conference on Computer Vision* (ECCV) (pp. 142–159). Springer.

[22] Zhuang, Z., & Zhou, Y. (2023). FasterVoxelPose+: Fast and accurate voxel-based 3D human pose estimation by depth-wise projection decay. In *Proceedings of the Asian Conference on Machine Learning* (ACML) (pp. 1763–1778). PMLR.

[23] Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 5686–5696). IEEE.

[24] He, Y., Yan, R., Fragkiadaki, K., & Yu, S.-I. (2020). Epipolar transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 7779–7788). IEEE.

[25] Xiang, D., Joo, H., & Sheikh, Y. (2019). Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR) (pp. 10965–10974). IEEE.

[26] Huang, C. Z. T., Jiang, S. A., Li, Y., & Zhang, Z. (2020). End-to-end dynamic matching network for multi-view multi-person 3D pose estimation. In *Proceedings of the European Conference on Computer Vision* (ECCV) (pp. 477–493). Springer.

[27] Zhu, Z., Liu, S., Shuai, J., & Zhu, W. (2023). 3D associative embedding: Multi-view 3D human pose estimation in crowded scenes. In *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things* (CNIOT) (pp. 131–139). ACM.