RePOSE: 3D Human Pose Estimation via Spatio-Temporal Depth Relational Consistency

Ziming Sun[1,2], Yuan Liang[1,2⋆], Zejun Ma[3], Tianle Zhang[1,2], Linchao Bao[4], Guiqing Li[1], and Shengfeng He[2(✉)]

[1] South China University of Technology, China
[2] Singapore Management University, Singapore
shengfenghe@smu.edu.sg
[3] Johns Hopkins University, USA
[4] Tencent AI Lab, China
https://github.com/StupidAdam/RePOSE

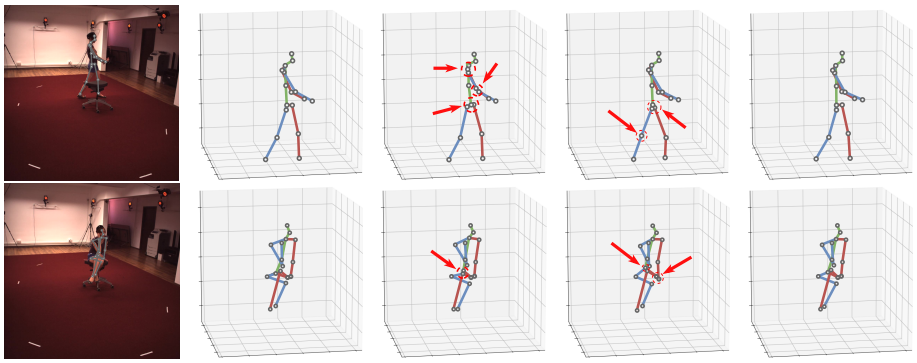
Abstract. We introduce RePOSE, a simple yet effective approach for addressing occlusion challenges in the learning of 3D human pose estimation (HPE) from videos. Conventional approaches typically employ absolute depth signals as supervision, which are adept at discernible keypoints but become less reliable when keypoints are occluded, resulting in vague and inconsistent learning trajectories for the neural network. RePOSE overcomes this limitation by introducing spatio-temporal relational depth consistency into the supervision signals. The core rationale of our method lies in prioritizing the precise sequencing of occluded keypoints. This is achieved by using a relative depth consistency loss that operates in both spatial and temporal domains. By doing so, RePOSE shifts the focus from learning absolute depth values, which can be misleading in occluded scenarios, to relative positioning, which provides a more robust and reliable cue for accurate pose estimation. This subtle yet crucial shift facilitates more consistent and accurate 3D HPE under occlusion conditions. The elegance of our core idea lies in its simplicity and ease of implementation, requiring only a few lines of code. Extensive experiments validate that RePOSE not only outperforms existing state-of-the-art methods but also significantly enhances the robustness and precision of 3D HPE in challenging occluded environments.

Keywords: 3D human pose estimation · depth relational loss functions

1 Introduction

The estimation of 3D human pose from videos is a foundational task in computer vision, with widespread implications across various applications such as virtual reality [4, 12, 21, 38], human-computer interaction [8, 36, 37], and action recognition [10, 34, 35, 40, 41]. This task involves localizing body joints in a three-dimensional space to create a comprehensive representation of the human body,

⋆ The first two authors contributed equally to this work.



(a) Input **(b)** GT **(c)** MotionBERT **(d)** PoseFormerV2 **(e)** Ours

Fig. 1: We introduce RePOSE, a novel solution designed to address the occlusion challenges in 3D pose estimation through relational supervision. Unlike existing methods (c) [46] and (d) [44], which struggle with occlusions due to their dependence on absolute depth signals, our approach leverages relational depth consistency, effectively overcoming these obstacles.

typically in the form of a skeletal model. The methodologies in 3D human pose estimation (HPE) are primarily categorized into two approaches: direct estimation and 2D-to-3D lifting methods. The former [22, 25] entails the direct estimation of human pose from provided videos, whereas the latter [2, 14, 17, 27, 42, 44–46] involves the initial estimation of intermediate 2D poses using off-the-shelf 2D detectors, followed by lifting them to 3D space. 2D-to-3D lifting methods have gained prominence due to the rapid evolution of 2D pose detection technologies, exhibiting superior performance in many scenarios.

Recent advancements in 2D-to-3D lifting methods often employ Transformer architectures [31], which have shown exceptional ability in capturing global information and modeling sequential data. These methods typically involve transforming detected 2D poses into 3D space using tailored algorithms [13–15,27,42,44–46]. However, this process inherently faces a many-to-one mapping challenge, where multiple distinct 3D poses can correspond to a single 2D projection, particularly under occlusion scenarios, as shown in Fig. 1c and 1d. In such cases, occlusion of key body joints leads to reduced accuracy in 2D pose estimations, creating ambiguities in the 3D reconstruction process. Moreover, relying on these unreliable occluded signals during model training results in vague and uncertain learning pathways. This limitation chiefly stems from the methods' reliance on absolute depth signals, which lose reliability under occluded conditions, thus undermining the overall accuracy of 3D pose estimation.

To address this challenge, we propose RePOSE, a novel approach specifically designed to enhance the accuracy of 3D HPE in occluded scenarios. RePOSE circumvents the limitations of absolute depth reliance by introducing a spatio-temporal relational depth consistency. Our method emphasizes the correct sequencing of occluded keypoints using a relative depth consistency loss in both

spatial and temporal domains, rather than depending solely on absolute depth values. This shift in approach enables our method to robustly handle occlusions, providing more accurate and reliable pose estimations, as shown in Fig. 1e. Additionally, the simplicity of spatio-temporal relational depth consistency, which can be implemented in just a few lines of code, makes it a practical and effective solution for real-world applications.

We demonstrate through extensive experiments on the Human3.6M [11] and MPI-INF-3DHP [20] benchmarks that our method outperforms state-of-the-art methods by a large margin, particularly excelling in scenarios involving occlusions. Our findings are further substantiated through visual comparisons and an ablation study, showcasing the effectiveness and innovation of RePOSE in tackling one of the most challenging aspects of 3D human pose estimation. This method lays the groundwork for future research aimed at enhancing pose estimation accuracy and reliability in complex environments.

2 Related Works

2D-to-3D Lifting HPE. The prevalent approach in 3D HPE involves lifting 2D pose sequences to 3D poses, a method enhanced by advanced 2D detectors like CPN [3], AlphaPose [7], and HRNet [33]. These detectors facilitate the conversion process with various techniques, including Fully Connected Networks (FCN) [6,19], Long Short-Term Memory (LSTM) [9,16,32], Graph Convolutional Networks (GCN) [1,5,39,43], Temporal Convolutional Networks (TCN) [2,17,26], and Transformers [13, 14, 27, 42, 44–46]. While SimpleBaseline [19] and LSTM-based methods [9] highlight the initial strides, they exhibit limitations in computational efficiency and temporal information utilization. GCN-based methods [1, 5, 39, 43], though adept at local joint modeling, struggle with computational complexity. Transformer-based architectures, such as PoseFormer [45], bring significant improvements by modeling spatial and temporal joint relations, albeit with limitations in learning spatial-temporal dependencies. The multi-solution nature of 2D-to-3D lifting is addressed by methods like MHFormer [14], which generate multiple hypotheses to account for ambiguity in body part locations. Our work challenges this approach, positing a single accurate solution corresponding to the original 2D pose, driven by spatial and temporal relational constraints to optimize results.

Loss Functions in 3D Human Pose Estimation. The accuracy of 3D HPE models heavily relies on the choice of loss functions. The Mean Per Joint Position Error (MPJPE) [14, 27, 44–46] is a prevalent metric, gauging the Euclidean distance between predicted and ground truth joint positions. Its simplicity and direct measurement of joint accuracy make it a popular choice. However, it does not account for joint orientations, potentially overlooking critical aspects of pose estimation. The Weighted-MPJPE (W-MPJPE) [42] attempts to refine this by assigning different weights to various joints, thereby prioritizing certain joints over others based on the application's needs. Yet, this method introduces its own set of challenges, including potential bias in weight assignment and the

need for expert knowledge in determining these weights. Additionally, the Mean Per Joint Velocity Error (MPJVE) [26,42] has been used to improve the temporal coherence between the predicted pose sequence and the ground truth sequence, which is a crucial aspect in dynamic pose estimation. However, it falls short in leveraging long-range temporal information. To address this gap, Zhang *et al.* [42] integrated the MPJPE and temporal consistency losses introduced by Hossain *et al.* [9] to leverage temporal relationships and smooth pose transitions. However, these adaptations insufficiently tackle the issue of relative depth perception among joints. Various studies [29,46] attempt to quantify limb angle inaccuracies by calculating the Mean Absolute Error between the estimated and ground-truth skeletal angles. Additionally, they explore angle dynamics through the Mean Per-Angle Velocity Error (MPAVE), representing the mean Euclidean distance in the angle's first temporal derivative. Despite these efforts, current methodologies overly focus on the x and y dimensions, derived primarily from 2D pose estimations, neglecting the critical depth (z-axis) disparities. Alternatively, Pavlakos *et al.* [25] utilize predicted depth values to compute the ordinal depth loss, especially when handling datasets that lack ground truth labels. However, this approach offers a weaker supervisory signal due to the inherent unreliability of the predictions.

In response to the above problems, we introduce two novel loss functions: spatial and temporal relational depth consistency losses. The spatial loss focuses on relative depth differences between joints within a single frame, ensuring accurate depth relationships and mitigating pose estimation inconsistencies. The temporal loss maintains pose smoothness over time, considering relative depth differences of the same joint across frames. This dual focus enhances the overall robustness of 3D HPE, ensuring more accurate, temporally coherent, and visually appealing human pose sequences. Our methodology, utilizing these losses, aims to significantly improve the accuracy and temporal stability of pose estimation, making it more applicable and reliable in various computer vision tasks.

3 RePOSE

3.1 Problem Formulation and Preliminaries

Our method leverages the 2D-to-3D lifting pipeline to elevate a sequence of 2D skeletons to their corresponding 3D coordinates. Given a sequence of 2D joints with confidence scores, represented as $X \in \mathbb{R}^{T\times J\times 3}$ where T and J denote the number of frames and joints respectively, this input is transformed into a d-dimensional feature space $F \in \mathbb{R}^{T\times J\times d}$ by the lifting network, with d varying across different network architectures. The output of this process is the estimated 3D pose sequence, denoted as $\hat{Y} \in \mathbb{R}^{T\times J\times 3}$, obtained through a regression mechanism. In our approach, we specifically treat the third dimension of $\hat{Y}$ as the depth feature $\hat{D} \in \mathbb{R}^{T\times J}$, facilitating the computation of relational depth consistency losses for model training supervision.

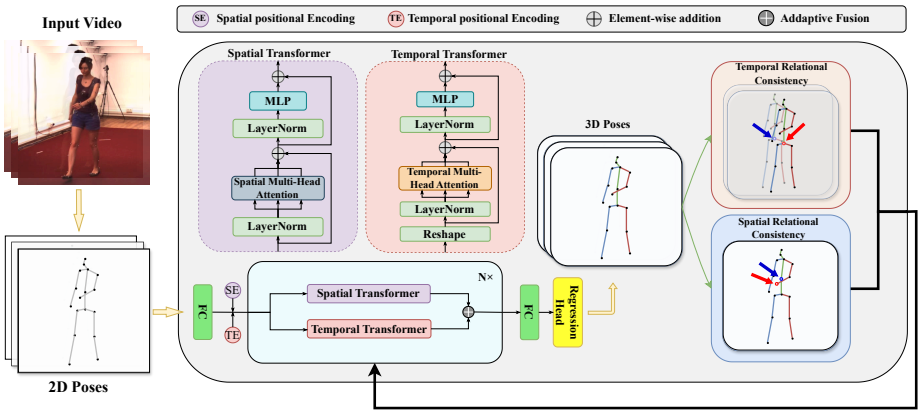


Fig. 2: Our framework pipeline. Our network is designed with a spatio-temporal architecture, enabling simultaneous modeling of both spatial and temporal information. The generated 3D poses are processed through dual loss modules: the spatial loss module focuses on the relative depth disparities among joints within each frame, while the temporal loss module evaluates the relative depth differences of identical joints across consecutive frames. This dual-module approach ensures comprehensive depth accuracy and temporal consistency in the pose estimation process.

Prevailing methodologies in recent studies [13,14,27,42,44–46] predominantly utilize Mean Per Joint Position Error (MPJPE) as loss function, defined as:

$$\mathcal{L}_{mpjpe} = \sum_{t=1}^{T} \sum_{i=1}^{J} \|\hat{y}_{t,i} - y_{t,i}\|_2, \tag{1}$$

where T and J represent the total number of frames and joints in each video, $\hat{y}_{t,i}$ and $y_{t,i}$ correspond to the estimated and ground-truth positions of keypoint i in frame t, respectively.

While the MPJPE is a widely accepted metric in 3D HPE, it has limitations, particularly in its focus on joint positions without explicit consideration of spatial relationships or pose topology. This becomes a significant concern in scenarios with occlusions, where understanding the relative positions of body parts is vital for accurate pose estimation. To address this, we introduce the spatial depth ranking loss, specifically designed to enhance pose topology awareness.

Additionally, conventional techniques often fail to effectively utilize temporal information, as they typically compute loss independently for each frame, neglecting the continuity and interdependence within video sequences. This oversight can significantly limit a model's ability to capture realistic human motion over time, a crucial aspect in dynamic pose estimation tasks. To mitigate this gap, we propose the temporal depth ranking loss, a novel component aimed at incorporating temporal dynamics for more accurate and lifelike pose estimation.

3.2 Network Structure

Inspired by [27,45], we recognize the exemplary performance of spatial-temporal transformer architectures in 3D human pose estimation (HPE) due to their robust capability to model both spatial and temporal information in video sequences. In particular, they demonstrate a parallel dual-stream design to be effective in ensuring comprehensive learning of spatial and temporal aspects. Consequently, our network, RePOSE, adopts a bifurcated structure of the spatio-temporal transformer to enhance the capture of spatial and temporal features from input videos.

As depicted in Fig. 2, RePOSE is composed of a spatio-temporal module and two specialized loss modules. The spatial transformer encoding module processes individual body joints as distinct tokens, thereby efficiently capturing the inter-joint relationships within the same frame. In contrast, the temporal transformer encoding module considers cross-frame features, enabling it to discern relationships between joints across different frames. The outputs from these modules are then directed to the spatial and temporal relational depth consistency modules, which compute the respective losses. A crucial aspect of our architecture is the backward pass, which plays a vital role in guiding the model's training process. The nuances and operations of the spatial and temporal relational depth consistency modules are elaborated in subsequent sections.

3.3 Spatial Relational Depth Consistency

In any given frame, the ideal relationship between estimated depths $\hat{d}$ and ground-truth depths d should adhere to the condition $\hat{d}_{t,i} < \hat{d}_{t,j}$ if $d_{t,i} < d_{t,j}$, where t denotes a specific time step, and $d_{t,i}$ and $d_{t,j}$ represent the depths of different joints i and j at that moment, respectively. The same applies for $\hat{d}_{t,i}$ and $\hat{d}_{t,j}$. However, we observe that this relative depth relationship is often overlooked in existing methods, leading to inaccurate results, especially in scenarios with noise such as self-occlusion. To address this, we introduce a novel loss function to the training procedure, aimed at steering the network optimization in a more accurate direction. The formula is as follows:

$$\mathcal{L}_S = \sum_{t=1}^{T} \sum_{d_{t,i}<d_{t,j}} max(0, \hat{d}_{t,i} - \hat{d}_{t,j}), \tag{2}$$

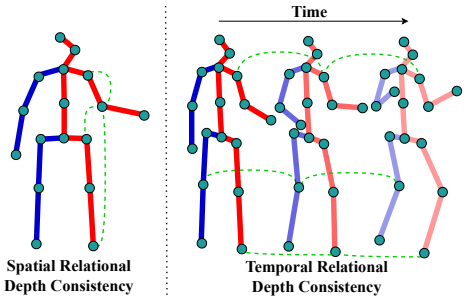


Fig. 3: Overview of spatio-temporal relational depth consistency loss. The spatial loss (left) enhances relative depth relationships between the in-frame joints. The temporal loss (right) ensures the consistent relative depth relationships of joints across frames. The green dashed lines illustrate examples where the loss is calculated for the relative depth of linked joints against the ground-truth relative depth.

Eq. (2) demonstrates that our model penalizes deviations in the ranking of estimated depths from the ground-truth rankings. An illustration of this loss is displayed in the left part of Fig. 3.

In cases where body parts are occluded, standard models often struggle to accurately determine the depth positioning of joints, leading to spatial inconsistencies. By integrating our spatial depth rank loss, we guide the estimations to align more closely with the actual spatial configuration, thereby providing feedback to the model for improved training and enhanced differentiation of obscured body parts or joints. Particularly in self-occlusion scenarios, traditional approaches might incorrectly associate body joints, such as mistaking a hand for an elbow due to overlapping positions. This type of error, which might go unnoticed when solely using MPJPE as the loss function, becomes apparent when the human skeleton is visualized. Our proposed spatial depth rank loss function, by reinforcing the inter-joint relationships within a frame, effectively prevents these errors and ensures a more accurate representation of the human body structure. The code of this loss is displayed in Algorithm. 1.

Algorithm 1 Spatial Relational Depth Consistency

```
def process(x, J, idx):
  x = x.unsqueeze(idx)
  return x.repeat_interleave(J, idx)
def loss_spatial(x, gt):
  #x, gt: [B, T, J, 3]
  _, _, J, _ = x.shape
  x = x[..., 2]
  gt = gt[..., 2]
  A = process(gt, J, -1)
  B = process(gt, J, -2)
  C = process(x, J, -1)
  D = process(x, J, -2)
  Comp_1 = (C - D) * (A < B)
  Comp_2 = torch.zeros_like(C)
  SRLoss = torch.maximum(
           Comp_1, Comp_2)
  return torch.mean(SRLoss)
```

Algorithm 2 Temporal Relational Depth Consistency

```
def process(x, T, idx):
  x = x.unsqueeze(idx)
  return x.repeat_interleave(T, idx)
def loss_temporal(x, gt):
  #x, gt: [B, T, J, 3]
  _, T, _, _ = x.shape
  x = x.permute(0, 2, 1)[..., 2]
  gt = gt.permute(0, 2, 1)[..., 2]
  A = process(gt, T, -1)
  B = process(gt, T, -2)
  C = process(x, T, -1)
  D = process(x, T, -2)
  Comp_1 = (C - D) * (A < B)
  Comp_2 = torch.zeros_like(C)
  TRLoss = torch.maximum(
           Comp_1, Comp_2)
  return torch.mean(TRLoss)
```

3.4 Temporal Relational Depth Consistency

In parallel with our spatial considerations, we address depth consistency in the temporal dimension. Temporally, the depth ranking of a specific joint should remain consistent across different time steps to ensure the continuity of the 3D pose sequence. However, challenges arise, particularly during self-occlusion, where the estimated poses may disrupt this continuity. To counteract this, we

introduce the following loss function:

$$\mathcal{L}_T = \sum_{j=1}^{J} \sum_{d_{m,j} < d_{n,j}} max(0, \hat{d}_{m,j} - \hat{d}_{n,j}), \tag{3}$$

where j indicates a specific joint, $d_{m,j}$ and $d_{n,j}$ represent the depths of joint j at different times m and n, similarly for $\hat{d}_{m,j}$ and $\hat{d}_{n,j}$. The right part of Fig. 3 shows the idea of this loss.

The temporal depth ranking loss encourages each joint in the estimated poses to maintain temporal consistency, preventing any joint from deviating suddenly from its original trajectory within the sequence. This loss is particularly crucial in ensuring that the temporal relationships between consecutive frames in video sequences are adequately maintained, which aids in smoothing the motion and making the model less susceptible to outliers. The implementation is detailed in Algorithm. 2.

3.5 Final Objective

Our comprehensive approach integrates the spatial and temporal dimensions to significantly enhance the coherence and realism of estimated 3D pose sequences. This integration ensures that the movements captured are not only spatially precise but also exhibit temporal fluidity and lifelike dynamics. Our final objective function combines the conventional MPJPE loss with our spatio-temporal relational depth consistency losses. The final loss function is formulated as follows:

$$\mathcal{L} = \mathcal{L}_{mpjpe} + \lambda_S \mathcal{L}_S + \lambda_T \mathcal{L}_T, \tag{4}$$

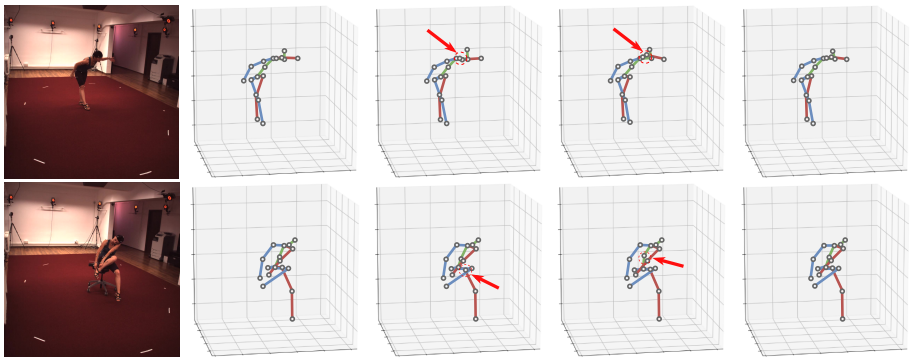
where $\mathcal{L}mpjpe$ is the standard MPJPE loss, and the parameters λ_S and λ_T are weighting factors that balance the contribution of the spatial and temporal components in the overall loss function. This unified loss function is designed to optimize the network in a holistic manner, addressing both the individual joint accuracy and the relational depth in both spatial and temporal dimensions.

4 Experiments

4.1 Datasets and Evaluation Metrics

We have conducted training and evaluation of our model using the datasets Human3.6M [11] and MPI-INF-3DHP [20].

Human3.6M is a well-established indoor dataset widely employed in the field of 3D HPE. This dataset comprises 3.6 million video frames captured from four distinct viewing angles, featuring performances by eleven professional actors engaged in diverse activities, including sitting, eating, and walking and more. In line with research [2, 17, 19, 26, 45, 46], we utilize subjects 1, 5, 6, 7, and 8 as our training dataset, reserving subjects 9 and 11 for testing purposes.



(a) Input (b) GT (c) MotionBERT (d) PoseFormerV2 (e) Ours

Fig. 4: Qualitative comparison on Human3.6M. Our method is evaluated against MotionBERT [46] and PoseFormerV2 [44] in scenarios featuring extensive self-occlusion of the human body. Areas with incorrect relative depth relationships are marked with red circles for emphasis, illustrating the effectiveness of each method in handling complex occlusion challenges.

In our evaluation process, we report two key metrics: MPJPE and Procrustes Aligned-MPJPE (PA-MPJPE), both measured in millimeters. MPJPE assesses the average Euclidean distance between the estimated joint positions and their corresponding ground truth values. In contrast, the PA-MPJPE represents the MPJPE values after performing a rigid alignment procedure involving translation, rotation, and scaling operations to bring the estimated 3D pose as close as possible to the ground truth poses.

MPI-INF-3DHP is another large-scale 3D HPE dataset consisting of both constrained indoor and challenging outdoor scenes. It records 8 actors performing 8 activities. It consists of over 1.3 million frames captured from the 14 cameras. To rigorously assess the generalization capabilities of RePOSE, particularly in challenging outdoor environments and scenarios involving occlusions, we conducted a comprehensive evaluation of our model using the MPI-INF-3DHP dataset. Here we utilize 2D pose sequences with a length of 81 frames each as input, as distinct from what it is in Human3.6M. Following previous works [2, 32, 44], we report MPJPE, Percentage of Correct Keypoints (PCK) within the 150mm range as well as Area Under the Curve (AUC).

4.2 Implementation Details

We have implemented our methodology using PyTorch [24] on a computing platform consisting of two NVIDIA RTX 3090 GPUs. Following previous studies [29, 44, 46], we have configured the video sequence length to 243 frames and leveraged the Stacked Hourglass [23] for the extraction of 2D poses during both training and evaluation phases. Additionally, we have applied horizontal flipping augmentation, consistent with [1, 2, 46]. For the optimization of model parameters, we have employed the AdamW [18] optimizer for 100 training epochs,

Table 1: Quantitative Comparison on Human3.6M Dataset. The top and bottom show comparison based on MPJPE and Procrustes-Aligned MPJPE in millimeters respectively. In this table, T represents the number of input frames used in each method. The best-performing methods are highlighted in bold, while the second-best results are underlined for clarity. Different 2D detection methods are indicated as follows: (∗) denotes the use of CPN [3], (†) indicates the use of HRNet [33], and (‡) signifies the use of SH [23] for 2D pose detection.

MPJPE	T	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
*VideoPose3D (CVPR'19) [26]	243	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
*PoseFormer (ICCV'21) [45]	81	41.5	44.8	39.8	42.5	46.5	51.6	42.1	42.0	53.3	60.7	45.5	43.3	46.1	31.8	32.2	44.3
*MHFormer (CVPR'22) [14]	351	39.2	43.1	40.1	40.9	44.9	51.2	40.6	41.3	53.5	60.3	43.7	41.1	43.8	29.8	30.6	43.0
†MixSTE (CVPR'22) [42]	243	36.7	39.0	36.5	39.4	40.2	44.9	39.8	36.9	47.9	54.8	39.6	37.8	39.3	29.7	30.6	39.8
*P-STMO (ECCV'22) [27]	243	38.4	42.1	39.8	40.2	45.2	48.9	40.4	38.3	53.8	57.3	43.9	41.6	42.2	29.3	29.3	42.1
*PoseFormerV2 (CVPR'23) [44]	243	41.2	45.4	41.4	44.0	46.6	53.7	42.6	42.6	55.1	64.6	45.7	42.8	45.7	32.3	32.8	45.2
*STCFormer (CVPR'23) [30]	243	38.4	41.2	36.8	38.0	42.7	50.5	38.7	38.2	52.5	56.8	41.8	38.4	40.2	26.2	27.7	40.5
*D3DP (ICCV'23) [28]	243	37.7	39.9	35.7	38.2	41.9	48.8	39.5	38.3	50.5	53.9	41.6	39.4	39.8	27.4	27.5	40.0
‡MotionBERT (ICCV'23) [46]	243	36.1	37.5	35.8	32.1	40.3	46.3	36.1	35.3	**46.9**	53.9	39.5	36.3	35.8	25.1	25.3	37.5
‡MoAGFormer (WACV'24) [29]	243	36.8	38.5	35.9	33.0	41.1	48.6	38.0	**34.8**	49.0	51.4	40.3	37.4	36.3	27.2	27.2	38.4
‡RePOSE (Ours)	243	**34.6**	**36.8**	**35.0**	**31.3**	**38.8**	**44.7**	**35.3**	35.5	47.5	**50.9**	**38.4**	**35.5**	**34.9**	**23.9**	**25.0**	**36.5**
PA-MPJPE	T	Dir.	Disc.	Eat.	Greet	Phone	Photo	Pose	Purch.	Sit	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg
*VideoPose3D (CVPR'19) [26]	243	34.1	36.1	34.4	37.2	36.4	42.2	34.4	33.6	45.0	52.5	37.4	33.8	37.8	25.6	27.3	36.5
*PoseFormer (ICCV'21) [45]	81	32.5	34.8	32.6	34.6	35.3	39.5	32.1	32.0	42.8	48.5	34.8	32.4	35.3	24.5	26.0	34.6
*MHFormer (CVPR'22) [14]	351	31.5	34.9	32.8	33.6	35.3	39.6	32.0	32.2	43.5	48.7	36.4	32.6	34.3	23.9	25.1	34.4
*MixSTE (CVPR'22) [42]	243	30.8	33.1	30.3	31.8	33.1	39.1	31.1	30.5	42.5	44.5	34.0	30.8	32.7	22.1	22.9	32.6
*P-STMO (ECCV'22) [27]	243	31.3	35.2	32.9	33.9	35.4	39.3	32.5	31.5	44.6	48.2	36.3	32.9	34.4	23.8	23.9	34.4
*PoseFormerV2 (CVPR'23) [44]	243	32.3	35.8	33.7	35.8	35.9	41.1	33.1	32.7	44.2	51.9	37.4	32.8	35.5	25.2	26.5	35.6
*STCFormer (CVPR'23) [30]	243	29.3	33.0	30.7	30.6	32.7	38.2	29.7	**28.8**	42.2	45.0	**33.3**	29.4	31.5	20.9	22.3	31.8
*D3DP (ICCV'23) [28]	243	30.6	32.5	**29.1**	31.0	**31.9**	37.6	30.3	29.4	40.6	**43.6**	**33.3**	30.5	31.4	21.5	22.4	31.7
‡MotionBERT (ICCV'23) [46]	243	29.5	31.3	32.0	27.0	32.8	37.2	29.8	29.6	41.0	48.2	34.3	30.1	30.2	21.2	21.8	31.7
‡MoAGFormer (WACV'24) [29]	243	31.0	32.6	31.0	27.9	34.0	38.7	31.5	30.0	41.4	45.4	34.8	30.8	31.3	22.8	23.2	32.5
‡RePOSE (Ours)	243	**29.1**	**30.9**	29.4	**26.6**	32.1	**35.8**	**28.6**	30.6	**39.9**	44.9	**33.3**	**29.0**	**29.6**	**20.1**	**21.2**	**30.7**

incorporating a weight decay factor of 0.01. The initial learning rate has been set to 5×10^{-4}, accompanied by a learning rate decay strategy with a decay rate of 0.99. Lastly, the hyper-parameters λ_S and λ_T are all set to 20. When training our model on Human3.6M, We utilize the stride data sample strategy with a one-third interval of the input length to make sure there are fewer overlapping frames between consecutive sequences. But for the MPI-INF-3DHP dataset, the interval is one-nine of the input length because of fewer samples.

4.3 Comparison with State-of-the-Art Methods

Human3.6M dataset. The proposed method has been seriously compared with state-of-the-art (SOTA) approaches, utilizing the Human3.6M dataset as the evaluation benchmark. As illustrated in Tab. 1, our method demonstrates exceptional performance, achieving an MPJPE of 36.5 millimeters and a PA-MPJPE of 30.7 millimeters. These results surpass those achieved by leading SOTA methods, indicating a notable performance improvement.

Table 2: Computational Complexity Comparison. The MPJPEs here are evaluated on Human3.6M using ground-truth 2D joint locations as input.

Method	Param (M)	MACs (G)	MPJPE
PoseFormer (ICCV'21) [45]	9.5	2.5	31.3
MHFormer (CVPR'22) [14]	30.9	7.0	30.5
MixSTE (CVPR'22) [42]	33.6	139.0	25.9
P-STMO (ECCV'22) [27]	6.2	0.7	29.3
PoseFormerV2 (CVPR'23) [44]	14.3	0.5	-
STCFormer (CVPR'23) [30]	4.7	19.6	21.3
MotionBERT (ICCV'23) [46]	42.5	174.7	17.8
MotionAGFormer (WACV'24) [29]	19.0	78.3	<u>17.3</u>
RePOSE (Ours)	16.0	43.7	**15.7**

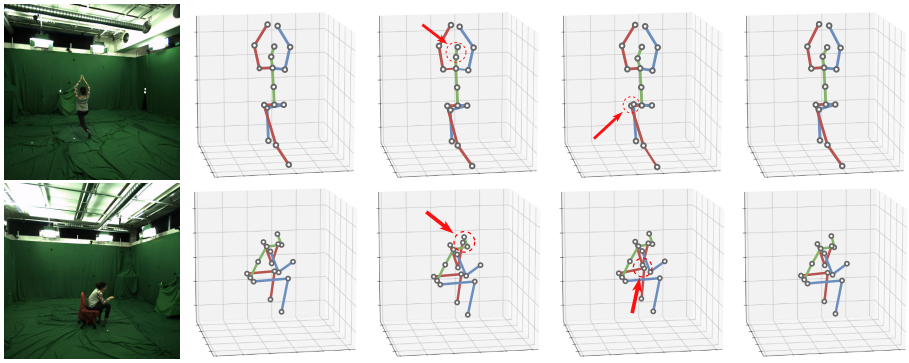
Table 3: Quantitative comparison on MPI-INF-3DHP dataset. The best and second-best results are bolded and underlined.

Method	PCK↑	AUC↑	MPJPE↓
VideoPose3D (CVPR'19) [26]	85.5	51.5	84.8
PoseFormer (ICCV'21) [45]	65.4	63.2	57.7
MHFormer (CVPR'22) [14]	93.8	63.3	58.0
MixSTE (CVPR'22) [42]	94.4	66.5	54.9
P-STMO (ECCV'22) [27]	97.9	75.8	32.2
D3DP (ICCV'23) [28]	97.7	77.8	30.2
PoseFormerV2 (CVPR'23) [44]	97.9	78.8	27.8
STCFormer (CVPR'23) [30]	**98.7**	83.9	23.1
MotionAGFormer (WACV'24) [29]	98.2	<u>85.3</u>	<u>16.2</u>
RePOSE (Ours)	<u>98.3</u>	**86.7**	**15.5**

It is worth noting that even when considering the best results reported by each of the prior studies, our method consistently outperforms them, especially in challenging scenarios involving actions such as "photo," "sitting," and "sitting down." This highlights its potential to effectively address challenges related to occlusion. Qualitative results can be found in Fig. 4.

We also compare the number of parameters, multiply–accumulate operations (MACs), and MPJPE using the ground-truth 2D pose sequences as input, as shown in Tab. 2. Our model attains the best MPJPE out of other works with limited parameters and algorithm complexity, demonstrating the effectiveness and efficiency of our proposed method.

MPI-INF-3DHP dataset. Tab. 3 presents a quantitative comparison of our method on the MPI-INF-3DHP dataset. RePOSE achieves impressive results, with an MPJPE of 15.5 mm and an Area Under Curve (AUC) of 86.7%, significantly outperforming other compared methods by margins of 1.8mm in MPJPE and 1.4% in AUC. However, in terms of the Percentage of Correct Keypoints (PCK), RePOSE reaches 98.3%, which is slightly lower by 0.4% compared



(a) Input (b) GT (c) D3DP (d) STCFormer (e) Ours

Fig. 5: Qualitative comparison on MPI-INF-3DHP. Our method is compared with D3DP [28] and STCFormer [30]. The areas with incorrect results are pointed out with red arrows and circles to exemplify the importance of relational depth consistency.

to the best performance achieved by existing methods, but we excel in the best balance of all metrics. For a more detailed visual assessment, refer to the qualitative comparisons illustrated in Fig. 5.

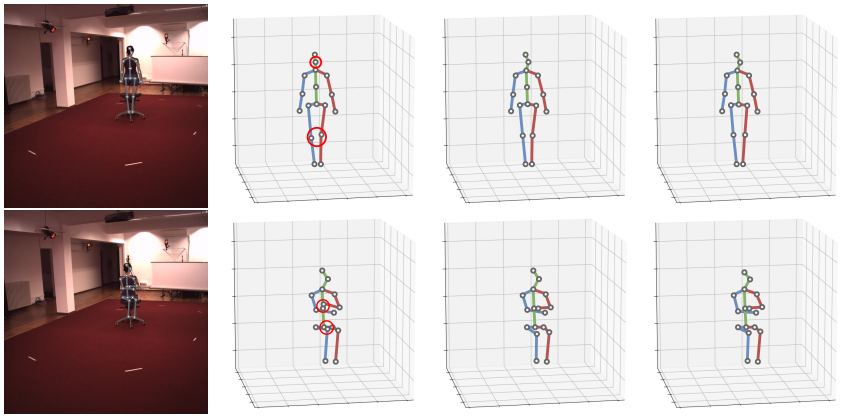
4.4 Ablation Study

To evaluate the distinct contributions of individual components, we incorporated spatial and temporal relational depth consistency loss functions separately into the baseline network. As shown in Tab. 4, it is clear that without these spatial and temporal depth constraints, the baseline model experiences degradation in both the MPJPE and PA-MPJPE metrics. However, the inclusion of either the spatial or temporal loss functions results in a significant enhancement in the model's performance. This observation underscores the effectiveness of these two loss functions in improving the overall quality of the model's predictions.

Table 4: Ablation Study on Individual Loss Functions. The study reveals that while both spatial and temporal loss functions individually contribute to improved pose estimation, their combined application yields the best results. Performance metrics have been evaluated on the Human3.6M dataset to substantiate this finding.

MPJPE	Spatial	Temporal	MPJPE↓	PA-MPJPE↓
✓	-	-	37.9	32.5
✓	✓	-	36.9	31.4
✓	-	✓	37.2	31.8
✓	✓	✓	**36.5**	**30.7**

Fig. 6 provides a comparative visualization of the performance of the RePOSE with and without the incorporation of two relational depth consistency losses. It is evident from the figure that the omission of these losses unfavorably affects the model's capacity to accurately estimate occluded body parts, as exemplified by the red circles. This inadequacy suggests that the model, without these losses, cannot effectively identify the relative depth of body parts during



(a) Input **(b)** w/o our losses **(c)** RePOSE **(d)** Ground Truth

Fig. 6: Visualized Results of Ablation Study. The second column displays poses generated by a model trained without the spatio-temporal relational depth consistency loss, contrasting with the third column, which shows results from the model including this loss. Joints with less accurate estimations are highlighted in red circles for clarity.

Table 5: Generalization evaluation on Human3.6M with our loss functions. With the integration of our loss functions, all methods exhibit improved performance. Δ indicates the difference between 'w/ST' and 'w/o'.

Method	T	MPJPE↓					PA-MPJPE↓				
		w/S	w/T	w/ST	w/o	Δ	w/S	w/T	w/ST	w/o	Δ
P-STMO-S [27]	27	45.3	45.5	**44.7**	46.1	-1.4	36.4	36.6	**35.6**	36.8	-1.2
STCFormer [30]	27	43.8	43.8	**43.3**	44.1	-0.8	34.1	34.0	**33.7**	34.8	-1.1
MotionBERT [46]	243	37.0	37.2	**36.7**	37.5	-0.8	31.4	31.7	**30.8**	31.8	-1.0

the training stage, leading to incorrect results. On the contrary, the integration of our proposed losses into the training phase significantly enhances the model's ability to resolve occlusions, thereby improving its estimation accuracy. This improvement is crucial for applications where precise depth perception is necessary, and it emphasizes the value of our contributions to the field of pose estimation in occlusion situations.

4.5 Generalization Evaluation

To validate the versatility and generalization capacity of our proposed loss functions, we extend their application to similar spatio-temporal transformer-based frameworks, including MotionBERT [46], STCFormer [30], and P-STMO [27]. We directly use their official pre-trained models with default setups and fine-tune them with our loss functions.

We detail the generalizability of our individual loss functions in Table 5. Overall, integrating both losses ('w/ST') leads to superior performance metrics compared to the original model configurations without our adjustments. Significantly, even when applied separately, each loss function ('w/S' or 'w/T') yields improvements across all evaluated methods, reinforcing their broad applicability. It is important to note that all reported results were achieved without specific tuning of the hyperparameters λ_S and λ_T. Therefore, there is an opportunity for further enhancements if these parameters were to be systematically optimized.

5 Conclusion and Discussion

In this paper, we introduce RePOSE, an approach designed to enhance 3D pose estimation accuracy in video sequences, specifically under conditions where body parts are obscured or occluded. Diverging from conventional methods that predominantly depend on depth information, which can falter when body parts are not visible, RePOSE adopts an innovative strategy centered on the relative positioning of body segments. This methodology not only exhibits superior efficacy in managing occlusions but also boasts a straightforward implementation, requiring minimal code adjustments. Our comprehensive experiments validate that RePOSE outperforms existing state-of-the-art techniques, particularly in scenarios marked by obstructions. These outcomes underscore RePOSE's efficacy and its promising potential in advancing 3D pose estimation, especially in practical settings where occlusions and partial visibilities are common. Additionally, the simplicity of its integration renders RePOSE a viable and effective solution for researchers aiming to refine pose estimation precision.

Limitations. Despite its significant advantages in occlusion scenarios within 3D human pose estimation, RePOSE encounters limitations, chiefly its dependence on the accuracy of preliminary 2D pose data. Discrepancies in this foundational data can adversely affect subsequent 3D estimations, a challenge inherent in 2D-to-3D conversion techniques. Moreover, although tailored for occlusions, RePOSE's performance may diminish in complex conditions, such as atypical body postures or interactions with objects, and in extremely low-light environments where depth cues are compromised.

Future Work. The efficacy of our newly proposed loss functions has been established across various spatio-temporal transformer architectures, including MotionBERT [46] and STCFormer [30]. Our next phase of research will explore the extension of these functions to a wider spectrum of models, including those based on GNNs and CNNs. Additionally, we aim to evolve the hyperparameters λ_S and λ_T from fixed values based on empirical selection to adaptive, learnable parameters optimized through the model's interaction with input data, thereby refining the model's depth dimension understanding in 3D human pose estimation. Future enhancements will focus on augmenting RePOSE's versatility across different input data forms and broadening its efficacy in a more extensive array of challenging conditions.

Acknowledgement

This project is supported by the Guangdong Basic and Applied Basic Research Foundation under project No. 2024A1515011995, Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020097, and National Research Foundation Singapore under the AI Singapore Programme under Grant AISG3-GV-2023-011.

References

1. Cai, Y., Ge, L., Liu, J., Cai, J., Cham, T.J., Yuan, J., Thalmann, N.M.: Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In: ICCV. pp. 2272–2281 (2019) 3, 9
2. Chen, T., Fang, C., Shen, X., Zhu, Y., Chen, Z., Luo, J.: Anatomy-aware 3d human pose estimation with bone-based pose decomposition. IEEE TCSVT **32**(1), 198–209 (2021) 2, 3, 8, 9
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR. pp. 7103–7112 (2018) 3, 10
4. Chen, Y., Tu, Z., Ge, L., Zhang, D., Chen, R., Yuan, J.: So-handnet: Self-organizing network for 3d hand pose estimation with semi-supervised learning. In: ICCV. pp. 6961–6970 (2019) 1
5. Ci, H., Wang, C., Ma, X., Wang, Y.: Optimizing network structure for 3d human pose estimation. In: ICCV. pp. 2262–2271 (2019) 3
6. Ci, H., Wu, M., Zhu, W., Ma, X., Dong, H., Zhong, F., Wang, Y.: Gfpose: Learning 3d human pose prior with gradient fields. In: CVPR. pp. 4800–4810 (2023) 3
7. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV. pp. 2334–2343 (2017) 3
8. Gong, J., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Meta agent teaming active learning for pose estimation. In: CVPR. pp. 11079–11089 (2022) 1
9. Hossain, M.R.I., Little, J.J.: Exploiting temporal information for 3d human pose estimation. In: ECCV. pp. 68–84 (2018) 3, 4
10. Huang, S., Wang, W., He, S., Lau, R.W.: Egocentric temporal action proposals. IEEE TIP **27**(2), 764–777 (2017) 1
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI **36**(7), 1325–1339 (2013) 3, 8
12. Jiang, Y., Zhou, Y., Liang, Y., Liu, W., Jiao, J., Quan, Y., He, S.: Diffuse3d: Wide-angle 3d photography via bilateral diffusion. In: ICCV. pp. 8998–9008 (2023) 1
13. Li, W., Liu, H., Ding, R., Liu, M., Wang, P., Yang, W.: Exploiting temporal contexts with strided transformer for 3d human pose estimation. IEEE Transactions on Multimedia **25**, 1282–1293 (2022) 2, 3, 5
14. Li, W., Liu, H., Tang, H., Wang, P., Van Gool, L.: Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In: CVPR. pp. 13147–13156 (2022) 2, 3, 5, 10, 11
15. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: CVPR. pp. 1954–1963 (2021) 2
16. Lin, M., Lin, L., Liang, X., Wang, K., Cheng, H.: Recurrent 3d pose sequence machines. In: CVPR. pp. 810–819 (2017) 3

17. Liu, R., Shen, J., Wang, H., Chen, C., Cheung, S.c., Asari, V.: Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In: CVPR. pp. 5064–5073 (2020) 2, 3, 8
18. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 9
19. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV. pp. 2640–2649 (2017) 3, 8
20. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 international conference on 3D vision (3DV). pp. 506–516. IEEE (2017) 3, 8
21. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM TOG **36**(4), 1–14 (2017) 1
22. Moon, G., Lee, K.M.: I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In: ECCV. pp. 752–768. Springer (2020) 2
23. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV. pp. 483–499. Springer (2016) 9, 10
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NeurIPS (2017) 9
25. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: CVPR. pp. 7307–7316 (2018) 2, 4
26. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR. pp. 7753–7762 (2019) 3, 4, 8, 10, 11
27. Shan, W., Liu, Z., Zhang, X., Wang, S., Ma, S., Gao, W.: P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In: ECCV. pp. 461–478. Springer (2022) 2, 3, 5, 6, 10, 11, 13
28. Shan, W., Liu, Z., Zhang, X., Wang, Z., Han, K., Wang, S., Ma, S., Gao, W.: Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In: ICCV. pp. 14761–14771 (October 2023) 10, 11, 12
29. Soroush Mehraban, Vida Adeli, B.T.: Motionagformer: Enhancing 3d human pose estimation with a transformer-gcnformer network. In: WACV (2024) 4, 9, 10, 11
30. Tang, Z., Qiu, Z., Hao, Y., Hong, R., Yao, T.: 3d human pose estimation with spatio-temporal criss-cross attention. In: CVPR. pp. 4790–4799 (2023) 10, 11, 12, 13, 14
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NeurIPS **30** (2017) 2
32. Wang, J., Yan, S., Xiong, Y., Lin, D.: Motion guided 3d pose estimation from videos. In: ECCV. pp. 764–780. Springer (2020) 3, 9
33. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE TPAMI **43**(10), 3349–3364 (2020) 3, 10
34. Xu, Y., Han, C., Qin, J., Xu, X., Han, G., He, S.: Transductive zero-shot action recognition via visually connected graph convolutional networks. IEEE TNNLS **32**(8), 3761–3769 (2020) 1
35. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: AAAI. vol. 32 (2018) 1

36. Yang, X., Xu, K., Chen, S., He, S., Yin, B.Y., Lau, R.: Active matting. NeurIPS **31** (2018) 1
37. Ye, M., Li, H., Du, B., Shen, J., Shao, L., Hoi, S.C.: Collaborative refining for person re-identification with label noise. IEEE TIP **31**, 379–391 (2021) 1
38. Yoon, J.S., Liu, L., Golyanik, V., Sarkar, K., Park, H.S., Theobalt, C.: Pose-guided human animation from a single image in the wild. In: CVPR. pp. 15039–15048 (2021) 1
39. Yu, B.X., Zhang, Z., Liu, Y., Zhong, S.h., Liu, Y., Chen, C.W.: Gla-gcn: Global-local adaptive graph convolutional network for 3d human pose estimation from monocular video. In: ICCV. pp. 8818–8829 (2023) 3
40. Zhang, C., Yang, T., Weng, J., Cao, M., Wang, J., Zou, Y.: Unsupervised pre-training for temporal action localization tasks. In: CVPR. pp. 14031–14041 (2022) 1
41. Zhang, J., Ye, G., Tu, Z., Qin, Y., Qin, Q., Zhang, J., Liu, J.: A spatial attentive and temporal dilated (satd) gcn for skeleton-based action recognition. CAAI Transactions on Intelligence Technology **7**(1), 46–55 (2022) 1
42. Zhang, J., Tu, Z., Yang, J., Chen, Y., Yuan, J.: Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In: CVPR. pp. 13232–13242 (2022) 2, 3, 4, 5, 10, 11
43. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: CVPR. pp. 3425–3435 (2019) 3
44. Zhao, Q., Zheng, C., Liu, M., Wang, P., Chen, C.: Poseformerv2: Exploring frequency domain for efficient and robust 3d human pose estimation. In: CVPR. pp. 8877–8886 (2023) 2, 3, 5, 9, 10, 11
45. Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., Ding, Z.: 3d human pose estimation with spatial and temporal transformers. In: ICCV. pp. 11656–11665 (2021) 2, 3, 5, 6, 8, 10, 11
46. Zhu, W., Ma, X., Liu, Z., Liu, L., Wu, W., Wang, Y.: Motionbert: A unified perspective on learning human motion representations. In: ICCV. pp. 15085–15099 (2023) 2, 3, 4, 5, 8, 9, 10, 11, 13, 14