Transformer-based Egocentric 3D Pose Estimation with Joint-Specific Weighted Average Pooling

1st Sungjin Hong
Contents Research Division, ETRI
Daejeon, Korea
sjhong0117@etri.re.kr

2nd Hye-sun Kim
Contents Research Division, ETRI
Daejeon, Korea
hsukim@etri.re.kr

3rd Cho-rong Yu
Contents Research Division, ETRI
Daejeon, Korea
crryu@etri.re.kr

***Abstract*—For egocentric 3D human pose estimation, we propose a Transformer-based framework with a Joint-Specific Weighted Average Pooling (JS-WAP) module that adaptively re-weights joint features, suppressing highly dynamic or occluded joints while emphasizing stable ones. By integrating stereo depth cues, JS-WAP enhances joint embeddings and improves the robustness of 3D regression. Experiments on a synthetic stereo egocentric dataset show that our method outperforms the UnrealEgo baseline, achieving 63.57mm MPJPE and 53.58mm PA-MPJPE, demonstrating accurate and robust performance.**

***Index Terms*—Egocentric 3D pose estimation, joint-specific weighted average pooling.**

I. Introduction

3D egocentric human pose estimation aims to reconstruct full-body 3D joint positions from bottom-view videos captured by body-mounted cameras, typically placed near the head. This technology has attracted growing interest due to its applications in inside-out tracking for VR/AR devices and immersive interaction. With the rapid expansion of wearable devices and the metaverse industry, there is an increasing demand for egocentric pose estimation systems that can robustly capture full-body motion without relying on external infrastructure.

Research in egocentric pose estimation has evolved along two main directions: single-camera and stereo-camera approaches. Single-camera methods, such as scene-aware techniques [1], leverage monocular RGB inputs and estimate surrounding depth information to alleviate physical inconsistencies (e.g., floating or penetrating limbs) caused by self-occlusion. While these approaches are lightweight and convenient, their accuracy is fundamentally limited by unreliable depth estimation from a single viewpoint. In contrast, stereo-camera methods directly exploit binocular disparity to recover depth cues, achieving more accurate 3D pose reconstruction.

Several representative works have advanced stereo-based egocentric pose estimation. EgoCap introduced a generative framework combined with a ConvNet-based body-part detector, showing robustness against severe self-occlusion, but the large 30–40 cm baseline between head-mounted cameras limited its practicality [2]. UnrealEgo proposed a large-scale synthetic dataset captured with eyeglasses-mounted stereo cameras and developed a weight-shared autoencoder with a dual-branch network for joint heatmap and 3D pose prediction [3]. However, its simple architecture did not explicitly address occlusion. More recently, 3D Human Pose Perception from Egocentric Stereo Videos leveraged a Transformer-based framework that integrates scene reconstruction and video-dependent query augmentation, significantly improving estimation accuracy, but the large network size and computational complexity raise concerns about efficiency, which may limit its applicability in certain real-world scenarios [4].

Despite these advances, severe self-occlusion remains a major challenge, and significant errors also occur in joints with large displacement, such as arms and legs. To address these limitations, we propose a Transformer-based egocentric 3D pose estimation framework with a novel Joint-Specific Weighted Average Pooling (JS-WAP) module, which adaptively suppresses unreliable joints while emphasizing stable ones, thereby improving the accuracy and robustness of 3D joint regression.

II. Proposed Framework for Egocentric 3D Pose Estimation

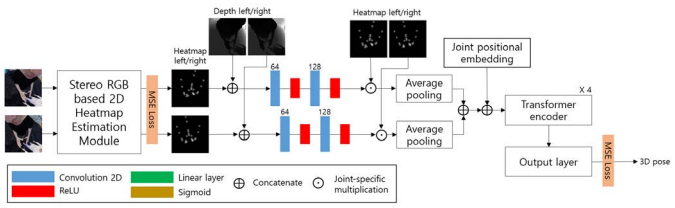

Fig. 1. Overview of the proposed egocentric 3D pose estimation framework. The system consists of a stereo 2D heatmap estimation module (adopted from UnrealEgo) and a 3D pose estimation module. The latter integrates depth-augmented joint features with the proposed joint-specific weighted average pooling (JS-WAP) and a Transformer encoder, producing accurate 3D joint coordinates.

The proposed framework for egocentric 3D pose estimation is illustrated in Fig. 1. The architecture is composed of two main modules: (1) a stereo heatmap estimation module that localizes 2D joints from stereo egocentric RGB inputs, and (2) a 3D pose estimation module that leverages the proposed joint-specific weighted average pooling (JS-WAP) in combination with a Transformer encoder to regress accurate 3D joint positions.

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2023-00224358, Multi-view wide sensing based XR high-DoF full body motion interface development)

TABLE I
COMPARISON OF MEAN PER JOINT POSITION ERROR (MPJPE, MM) FOR ARMS/LEGS/WHOLE AND PROCRUSTES-ALIGNED MPJPE (PA-MPJPE, MM) FOR WHOLE BODY.

Method	MPJPE (Arms)	MPJPE (Legs)	MPJPE ↓ (Whole)	PA-MPJPE ↓ (Whole)
$UnrealEgo$	100.49	107.19	82.54	68.95
$Ours$ RGB, w/o JS-WAP	89.64	90.24	71.24	59.83
$Ours$ RGBD, w/o JS-WAP	81.48	85.18	66.92	55.71
$Ours$ RGB, w/ JS-WAP	88.99	87.33	69.31	56.58
$Ours$ RGBD, w/ JS-WAP	73.50	80.61	63.57	53.58

A. Stereo RGB based 2D Heatmap Estimation Module

We adopt the stereo 2D heatmap estimation network introduced in UnrealEgo as our front-end. Given paired stereo RGB inputs, this module predicts per-joint 2D heatmaps in both left and right views. The network is trained with an MSE (mean squared error) loss against ground-truth 2D heatmaps [2].

B. Joint-Specific Weighted Average Pooling with Depth-Augmented Features

After obtaining stereo heatmaps from the UnrealEgo backbone, we enrich the joint representation with geometric depth cues from stereo depth. Each depth map is concatenated with the corresponding 2D heatmap, and the fused signals are passed through a set of convolutional layers with ReLU activations. This process yields joint-aware feature embeddings that integrate both appearance and depth information.

While UnrealEgo provides robust 2D heatmap estimation, our design ensures that the subsequent 3D reasoning is guided by richer joint descriptors. However, directly averaging joint embeddings treats all joints equally, regardless of their reliability under egocentric conditions. To address this, we introduce Joint-Specific Weighted Average Pooling (JS-WAP).

Concretely, for each joint j, its feature embedding f_j, is scaled by a learnable weight w_j producing a re-weighted representation $\tilde{f}_j = w_j \cdot f_j$. These joint-specific embeddings are then aggregated via weighted average pooling to form a compact global representation:

$$F = \frac{1}{N} \sum_{j=1}^{N} \tilde{f}_j$$

where N denotes the number of joints. This mechanism allows the model to suppress unreliable or occluded joints (e.g., feet hidden by the torso) while emphasizing stable and consistently visible ones (e.g., head, shoulders). By guiding the Transformer with structurally refined joint tokens, JS-WAP improves robustness against occlusion and enhances the accuracy of 3D pose regression. The network is trained using an MSE loss.

III. EXPERIMENT & RESULT

For training and evaluation, we constructed a synthetic egocentric dataset consisting of synchronized stereo RGB and stereo depth, together with ground-truth 2D and 3D annotations for 22 body joints: Head, neck pelvis, spine(low/mid/upper), arms(L/R shoulder, elbow, hand, finger tip), legs(L/R thigh, knee, foot, toe). Data were collected in diverse environments, specifically office, parking lot, street, kitchen, and living room scenes, to capture a wide range of visual conditions and occlusion patterns. For subject diversity, we employed eight virtual human characters (four male and four female) with different body shapes and appearances. The dataset comprises a total of 15,580 samples, among which 12,464 are used for training and 3,116 are reserved for testing.

The training procedure consists of three stages. We first train the stereo heatmap estimation module with 2D keypoint supervision following UnrealEgo. Then, the proposed depth-augmented JS-WAP module is trained to learn joint embeddings. Finally, all components are integrated and fine-tuned end-to-end using an MSE loss on 3D joint coordinates. We adopt the Adam optimizer(1×10^{-4}) and train the network for 500 iterations.

As shown in Table 1, we compare the performance of our proposed method with the UnrealEgo baseline under different input modalities (RGB vs. RGBD) and with or without the JS-WAP module. Overall, our Transformer-based framework consistently outperforms UnrealEgo, demonstrating the benefit of Transformer-based modeling for egocentric 3D pose estimation. Incorporating stereo depth further enhances accuracy: for example, RGBD without JS-WAP achieves 66.92 MPJPE / 55.71 PA-MPJPE compared to 71.24 / 59.83 with RGB only. A finer analysis by body parts confirms this trend, as RGBD input reduces arm error to 81.48 (vs. 89.64 with RGB) and maintains competitive performance on the legs (85.18 vs. 90.24), showing the advantage of stereo depth cues in capturing more accurate geometric representations.

Introducing the JS-WAP module further improves stability and robustness against challenging joints with high motion variation or frequent occlusion. For instance, RGB with JS-WAP improves the RGB baseline to 69.31 / 56.58 overall, while combining RGBD with JS-WAP yields the most significant gains: arm error drops from 81.48 to 73.50, leg error from 85.18 to 80.61, and overall error reaches 63.57 MPJPE / 53.58 PA-MPJPE. These results highlight that JS-WAP is particularly effective in handling occlusion-prone and highly dynamic limb joints, validating three key insights: (1) Transformer-based architectures surpass the prior baseline, (2) stereo depth cues significantly enhance pose estimation accuracy, and (3) JS-WAP substantially improves robustness, especially for arms and legs.

REFERENCES

[1] Wang, Jian, et al. "Scene-aware egocentric 3d human pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[2] Rhodin, Helge, et al. "Egocap: egocentric marker-less motion capture with two fisheye cameras." ACM Transactions on Graphics (TOG) 35.6 (2016): 1-11.

[3] Akada, Hiroyasu, et al. "Unrealego: A new dataset for robust egocentric 3d human motion capture." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

[4] Akada, Hiroyasu, et al. "3d human pose perception from egocentric stereo videos." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2024.