EgoCoord: Self-calibrated Egocentric 3D Body Pose Estimation using Pixel-wise Coordinate Encoding

Jong-Bae Lee[1], Hyoung Lee[1], Beom-Ryeol Lee[1],
Byung-Gook Lee[2], and Wook-Ho Son[1,⋆]

[1] Electronics and Telecommunications Research Institute, Republic of Korea
{ttdlwhd, hyoung0708, lbr, whson}@etri.re.kr
[2] Department of Computer Engineering, Dongseo University, Republic of Korea
lbg@dongseo.ac.kr

Abstract. The primary challenges for egocentric 3D human pose estimation techniques are the perspective and radial distortions introduced by fisheye lenses. Previous methods utilized camera calibration for undistortion or utilized neural networks to regress 3D human poses from distorted 2D poses. In this paper, we propose a novel approach that integrates a pixel-wise coordinate encoding technique for recognizing image distortion and utilizes the Vision Transformer (ViT) to extract distortion and pose tokens from the input image. The extracted tokens are used in a 3D volumetric heatmap-based egocentric pose estimator, which predicts the 3D human pose using pose tokens and performs pose correction using distortion tokens. The approach integrates CoordConv's positional encoding strategies, neural network-based camera calibration methods, and the volumetric heatmap-based 3D human pose estimation method. We evaluate the proposed model's performance using our new evaluation dataset and compare it with state-of-the-art models. Additionally, we perform an ablation study to demonstrate the individual effects of each module in the proposed model.

Keywords: Egocentric 3D body pose estimation · Auto-calibration · Fisheye camera

1 Introduction

Egocentric 3D body pose estimation aims to determine the positions of each joint in body images captured by cameras mounted on the head. This technique enables for the estimation of a user's posture using only the cameras attached to head-mounted displays (HMDs), making it a vital area of applied research for future virtual reality (VR) and augmented reality (AR) environments. Recent improvements in mobile hardware and deep neural networks have not only made pose estimation from monocular RGB images more accurate, but they

[⋆] Corresponding authors

have also made it facilitated the emergence of accurate stereo RGB image-based pose estimation [1,2,15]. This paper focuses on pose estimation using monocular RGB images. Egocentric 3D body pose estimation commonly employs fisheye lens cameras that provide a 180-degree field-of-view (FOV) to capture a wide area. Consequently, images used in egocentric 3D body pose estimation exhibit stronger distortions compared to those captured by external cameras. These distortions result in higher errors in egocentric 3D body pose estimation compared to standard 3D body pose estimation. Therefore, the primary challenge in this field is correcting these distortions, with much of the research focused on developing methods to mitigate these effects. Egocentric 3D body pose estimation typically involves two steps: 1) feature extraction from the input RGB image; and 2) 3D pose estimation using the extracted features. While similar to external camera-based pose estimation, the main difference lies in handling image distortions. Common methods include 2D-3D projection, which uses a fisheye camera model [25]; 2D-3D lifting, which estimates 3D poses directly from 2D images without depth or a camera model [2,20,21]; and volumetric heatmap regression, which estimates 3D poses as 3D heatmaps without first estimating 2D poses. Recent studies have used volumetric heatmap regression, estimating a volumetric heatmap for the UV coordinates and depth in image space and subsequently transforming these into 3D space using a fisheye camera model [24]. Methods that rely on fisheye camera models typically depend on precise calibration data to achieve high accuracy. However, when trained on datasets with inaccurate calibration, these methods suffer from performance degradation compared to training on a well-calibrated dataset. This issue is further exacerbated during pretraining on multiple datasets, where calibration data may be incomplete or inaccurate. To address this limitation, we propose a method that removes the dependency on camera calibration data, thereby enabling the use of a wider variety of datasets and improving the robustness of pose estimation. In this research, we introduce a pixel-wise positional encoding method to correct the distortions introduced by fisheye lenses. We also present CoordViT, a transformer network that incorporates this encoding method. Further, we propose a volumetric heatmap-based 3D pose estimator that utilizes the features extracted from CoordViT to estimate undistorted 3D poses.

Our primary contributions are summarized as follows:

- We propose an end-to-end network that simplifies the pipeline from feature extraction to pose estimation, avoiding the need for complex configurations traditionally required to address image distortions.
- We introduce a novel encoding strategy for modeling image distortions directly within a neural network and propose a backbone network that utilizes this strategy.
- We present a volumetric heatmap-based 3D pose estimator that corrects distortions through self-calibration using a distortion parameter estimation network, thereby enhancing flexibility and applicability in diverse settings.

2 Related Works

Third-person-view 3D Body Pose Estimation. Third-person-view 3D body pose estimation is a widely researched area that involves estimating body postures using images captured by external cameras, often including multiple individuals in a scene. This contrasts with egocentric pose estimation, where the focus is usually on a single individual. In third-person-view estimation, two main approaches are employed to handle multiple subjects: the top-down method, which first detects individuals in the image and then estimates their poses, and the bottom-up method, which first identifies all joint keypoints and then groups them into individual poses. Early research focused on directly regressing 3D joint coordinates from extracted features [17, 27]. However, with the advancement of 2D pose estimation techniques, subsequent methods leveraged accurate 2D pose estimates to regress from 2D to 3D [4, 12, 19], improving generalization performance in real-world settings. Since 2D poses lack depth information, various techniques have been developed to address this limitation [4, 19]. Volumetric-based methods further improved 3D pose estimation accuracy by using 3D heatmap representations instead of directly regressing joint positions [11, 13, 14, 18]. Recent approaches have also explored the use of parameterized human models, such as SMPL [10], allowing for regression on model parameters in techniques like Human Mesh Recovery (HMR) [5–7].

Egocentric 3D Body Pose Estimation. Egocentric 3D body pose estimation uses head-mounted cameras to capture images of the wearer's body, offering an alternative to IMU-based motion capture systems by reducing spatial constraints. These systems often rely on fisheye lenses with a wide field of view (FOV), which introduce distortions that conventional third-person-view methods struggle to handle. Addressing these distortions is a key challenge in this field, leading to the development of fisheye camera models designed to account for lens distortions [16]. Research is focused on integrating these models into neural networks to minimize 3D errors. There are two main approaches in egocentric pose estimation: those that use camera models and those that do not. Methods without camera models directly estimate undistorted 3D poses from features, using neural networks to transform 2D poses from distorted images into 3D poses (2D-3D lifting) [2, 9, 20, 21]. On the other hand, methods using camera models predict 2D poses and depth information from distorted images and then apply a fisheye camera model to estimate the 3D pose. Recent research has explored predicting 2D heatmaps and depth information simultaneously, with volumetric heatmap representations being the latest development [24]. While methods without camera models can estimate poses without calibration data, their accuracy may drop if the distortion characteristics of the target images differ from those in training. Conversely, methods using camera models can achieve higher accuracy but are more dependent on accurate calibration data, making pose estimation difficult when such data is unavailable.

Auto-calibration on 3D Body Pose Estimation. Transforming between 2D and 3D poses requires a camera model and corresponding parameters, which define the relationship between image plane coordinates and world coordinates. Camera parameters are categorized as intrinsic, related to the camera's optical properties, and extrinsic, describing the transformation between world space and camera space. Traditionally, camera calibration is performed using images with calibration patterns. However, recent research has introduced neural networks that infer camera parameters directly from input images. In parametric human models like SMPL [10], extrinsic camera parameters are estimated to define the spatial relationship between the camera and the subject. In egocentric 3D body pose estimation, neural networks now predict intrinsic and distortion parameters from images, allowing for automatic distortion correction and reducing pose estimation errors without manual calibration [26].

3 Proposed Method

This section outlines a methodology designed to accommodate the radial distortion inherent in fisheye lenses during egocentric 3D body pose estimation. Our proposed method is illustrated in Fig. 1.

3.1 Coord Vision Transformer

This methodology integrates the radial distortion characteristics of fisheye lenses into transformer networks by leveraging the CoordConv [8] approach, which provides pixel-wise positional encoding within convolutional networks to enhance object detection and localization tasks. Conventionally, transformer networks process input tokens in parallel rather than sequentially and therefore do not intrinsically manage the sequential order of input data. In natural language processing, this challenge is addressed by applying positional encodings to each token, enabling the network to incorporate the spatial information of input tokens. However, in Vision Transformers, these input tokens correspond to segmented image patches and lack granular positional data for individual pixels. Additionally, most Vision Transformers use either sinusoidal positional encodings or learnable positional embeddings. While sinusoidal positional encodings effectively represent the absolute positions of input tokens, their periodic signal representation precludes linear processing of radial distortions. According to Scaramuzza's fisheye camera model [16], radial distortions are articulated as polynomials of radial distance and coefficients. To adapt this model for our network, we augment the input image with pixel-wise positional encoding. CoordConv originally proposed appending a channel-wise scaled [-1, 1] Cartesian coordinate representation (x, y) to the image. For our proposal, addressing radial distortions requires the radial distance from the image; thus, we append a scaled [0, 1] representation of the polar coordinate system's radial coordinate (r) channel-wise to the image. This representation yields a value of 0 at the

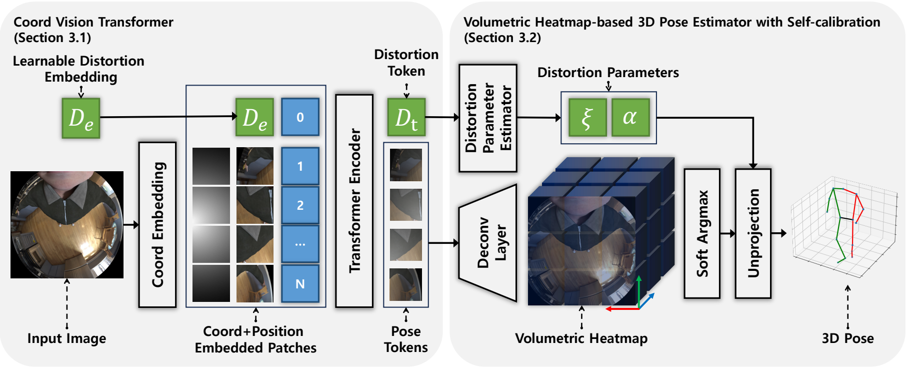

Coord Vision Transformer (Section 3.1)
Learnable Distortion Embedding
D_e
Coord Embedding
D_e
0
1
2
...
N
Transformer Encoder
Distortion Token
D_t
Input Image
Coord+Position Embedded Patches
Pose Tokens
Volumetric Heatmap-based 3D Pose Estimator with Self-calibration (Section 3.2)
Distortion Parameter Estimator
Distortion Parameters
ξ
α
Deconv Layer
Soft Argmax
Unprojection
Volumetric Heatmap
3D Pose

Fig. 1: An overview of our egocentric 3D body pose estimation pipeline. Our pipeline consists of two networks; the Coord Vision Transformer (CoordViT) and a volumetric heatmap-based 3D pose estimator. CoordViT is a transformer network that integrates an encoding technique specifically designed to extract distortion information from fisheye images (see Sec. 3.1 for details). The volumetric heatmap-based 3D pose estimator utilizes the feature tokens extracted by CoordViT to estimate the corrected 3D pose (see Sec. 3.2 for details).

image's principal point and 1 at its peripheral extremity. Following the application of pixel-wise positional encoding, the transformer encoder then extracts the distortion and pose tokens. These processes are shown in Fig. 2. To handle distortion information, we add a distortion embedding at the beginning of the image patches. This embedding is extracted as a distortion token in CoordViT. We use the distortion token to predict the distortion parameters of the input image, which we then use to correct the distortion in the 3D pose. During training, the self-attention mechanism captures features related to distortion and extracts the distortion token accordingly. These processes are illustrated in the Coord Vision Transformer part of Fig. 1.

In this study, we employ fine-tuning on a pretrained ViTPose to train CoordViT. However, due to the differing network parameters between ViTPose and CoordViT, a transformation process for these parameters is required prior to fine-tuning. The primary difference in parameters between CoordViT and ViTPose lies in the patch embedding at the transformer encoder's input stage. While ViTPose processes a 3-channel RGB image, CoordViT is designed to handle a 4-channel RGB+R image, which incorporates pixel-wise positional encoding. Patch embeddings are implemented through linear projection, resulting in network parameters represented as a [3 x 786] matrix for ViTPose and a [4 x 786] matrix for CoordViT. The procedure for initializing ViTPose's parameters for use in CoordViT involves two steps: First, a [4 x 786] matrix is initialized with a uniform distribution with a mean of 0 and a standard deviation of 0.01. Second, the first three rows of this matrix, [3 x 786], are replaced with the parameter values from ViTPose. Although the weights corresponding to pixel-wise posi-

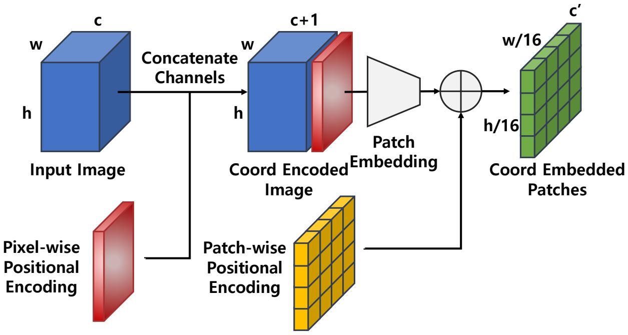


Fig. 2: Conceptual outline of Coord Embedding module in CoordViT network. Pixel-wise Positional Encoding provides coordinate information for image pixels, Patch-wise Positional Encoding provides positional information for patches of the image.

tional encoding initially have minimal values, implying a limited impact, they are subject to gradual modification during the fine-tuning process.

3.2 Volumetric Heatmap-based 3D Pose Estimator with Self-calibration

From the tokens extracted by the Coord Vision Transformer (CoordViT), the distortion token is employed for distortion parameter estimation. In the proposed network, we utilize the Double Sphere camera model instead of Scaramuzza's fisheye camera model [16], which relies on polynomial parameters. The polynomial parameters in Scaramuzza's model require the estimation of numerous parameters for accurate results. Moreover, the polynomial parameters are highly sensitive to small variations, which can lead to significant changes in the results. Additionally, the large number of parameters requires an even more complex neural network for accurate results. Consequently, we adopted the Double Sphere camera model [22], which represents distortion parameters with only two values, α and ξ, within predefined parameter ranges. This facilitates direct regression through the application of activation functions in neural networks. The distortion token undergoes linear projection to estimate the distortion parameters α and ξ, with α using a sigmoid activation function and ξ using a ReLU activation function to derive the final distortion parameters. The distortion parameter estimator is trained in an end-to-end manner to find the camera parameters that minimize the reprojection error of the final predicted 3D pose for the input image. Consequently, the model can be trained using only the input images and 3D poses, without requiring ground truth camera parameters. The features extracted by CoordViT are used to predict the 3D body pose using a volumetric heatmap-based pose estimation network. The proposed pose estimation network estimates volumetric heatmaps, which are then converted into 3D coordinates using a soft-argmax function. The soft-argmax function for converting the volumetric heatmap $\mathbf{HM}_j$ of joint j into 3D pose $\mathbf{P}_j$ of joint j can be written as:

$$\mathbf{P}_j = \sum_{x=0}^{W}\sum_{y=0}^{H}\sum_{z=0}^{D} \text{softmax}(\mathbf{HM}_j)_{x,y,z} \tag{1}$$

The 3D body pose obtained through the soft-argmax function corresponds to the fisheye camera space. Since the fisheye camera space represents distorted pose information, it is necessary to transform this into world space to obtain the final pose estimation. This transformation uses the Double Sphere camera model [22] and distortion parameters α and ξ. The unprojection function $\pi^{-1}(u, v, d)$ for converting from fisheye camera space $[u, v, d]^T$ to world space $[x, y, z]^T$ is as follows:

$$[x, y, z]^T = d\frac{D}{|D|} \tag{2}$$

$$D = \frac{m_z\xi + \sqrt{m_z^2 + (1-\xi^2)r^2}}{m_z^2 + r^2}\begin{bmatrix} m_x \\ m_y \\ m_z \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \xi \end{bmatrix} \tag{3}$$

$$m_x = \frac{u - c_x}{f_x} \tag{4}$$

$$m_y = \frac{v - c_y}{f_y} \tag{5}$$

$$r^2 = m_x^2 + m_y^2 \tag{6}$$

$$m_z = \frac{1 - \alpha^2 r^2}{\alpha\sqrt{1 - (2\alpha - 1)r^2 + 1 - \alpha}} \tag{7}$$

where ξ and α are estimated distortion parameters. In the field of egocentric pose estimation, the camera coordinate system is predominantly used as the frame of reference for 3D poses. This coordinate system uses the camera's sensor location as the origin, which means that each joint value of a 3D pose is represented relative to the camera's position. However, variations in camera extrinsic parameters, such as the wearer's height, can lead to estimation errors. This occurs because, despite changes in depth, the pose displayed in the fisheye camera image remains unchanged. To address this scaling issue, we use normalized camera coordinates instead of direct camera coordinates. This approach effectively mitigates the decrease in estimation accuracy caused by changes in scale.

The proposed method offers several advantages over the 2D-3D lifting method. First, it enhances the learning of 3D features by the feature extractor. The 2D-3D lifting method, which estimates 3D poses from 2D poses, involves separate networks for 2D and 3D estimations. Consequently, the 3D pose estimator only receives information from the 2D pose, guiding the feature extractor to prioritize enhancing the accuracy of the 2D pose estimator. In contrast, the volumetric heatmap-based method uses all features extracted by the feature extractor for estimating 3D poses, thus orienting the learning process towards improving the accuracy of the 3D pose estimator. This structure enables a simplified network configuration capable of handling radial distortions end-to-end without the need for additional network branches for estimating distortion coefficients or joint depths, as required in previous studies. Second, the use of volumetric heatmaps addresses issues related to scale differences. Unlike 2D heatmaps, which lack depth information and can lead to accuracy issues due to variations in the extrinsic parameters of the input images, volumetric heatmaps overcome this limitation by incorporating scale variance directly into their structure.

4 Experiment and Results

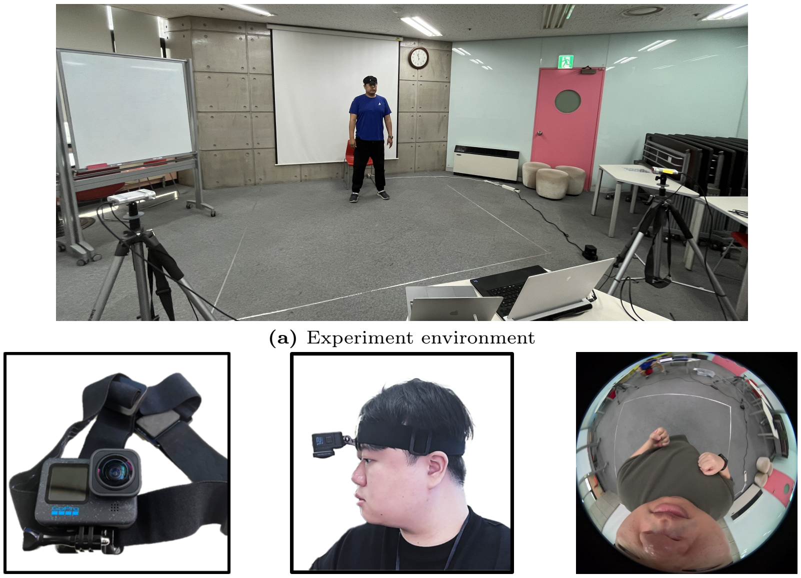
4.1 Egocentric Real-Life Dataset

In our experiment, we created a new evaluation dataset simulating HMD usage. Existing datasets use controlled settings, where the fisheye camera is positioned about 7cm away from the head to capture the full body, which differs from its actual position when mounted on an HMD. To address this, we set up an egocentric video capture environment using consumer-grade equipment, including a GoPro 12 Black with a fisheye lens and a head strap. This setup captured videos at a 177° FOV and a resolution of 3840×3360 at 60 fps. The recording was conducted in typical office/lounge environments without chroma key screens to simulate natural settings. We captured 45K frames from four actors (3M/1F) performing everyday activities such as stretching, walking, and sitting, with each session lasting 5 minutes. After each session, the head strap was reattached to slightly change the camera position. For 3D pose capture, we used two Azure Kinect DKs, and the estimated 3D poses from the depth images were used as pseudo-ground truth.

4.2 Experimental Setup

Training and Evaluation Dataset. Our proposed network was trained using the EgoWholeBody dataset [24]. The EgoWholeBody is a synthetic dataset that comprises 700K frames, captured from 14 characters, and includes ground truth annotations for 3D body poses. To evaluate the performance of the proposed network, we used the SceneEgo test dataset [23] and our proposed new dataset. SceneEgo consists of 28K frames recorded with two actors. To evaluate each dataset, we fine-tuned the network using the training set data provided by each dataset and then proceed with the evaluation of the network.

Evaluation Metrics. To evaluate the accuracy of egocentric 3D body pose estimation, we employ two metrics: mean per joint position error (MPJPE) and procrustes-aligned MPJPE (PA-MPJPE). MPJPE calculates the average euclidean distance between the predicted pose values and the ground truth. PA-MPJPE, on the other hand, evaluates the predictions after applying rigid alignment to the ground truth using procrustes analysis. The SceneEgo dataset was



(a) Experiment environment

(b) Fisheye camera

(c) Fisheye camera setup

(d) Egocentric view

Fig. 3: The setup of the egocentric fisheye camera and one example of the egocentric images.

evaluated using both MPJPE and PA-MPJPE, whereas our proposed dataset was evaluated only using PA-MPJPE. MPJPE of the predicted body pose $\hat{\mathbf{P}}$ and the ground truth body pose $\mathbf{P}$ can be written as:

$$E(\mathbf{P}, \hat{\mathbf{P}}) = \frac{1}{N_f} \frac{1}{N_j} \sum_{f=1}^{N_f} \sum_{j=1}^{N_j} ||\mathbf{P}_j^f - \hat{\mathbf{P}}_j^f||_2 \tag{8}$$

where $\mathbf{P}_j^f$ is the j-th joint position of the f-th frame.

Implementation Details. The weights of CoordViT were initialized using the pretrained weights from the ViTPose-B network. Due to the different parameter shapes used in CoordViT's patch embedding linear projection layer, the following initialization methods were employed: 1) The parameters of CoordViT's linear projection layer [4 x 786] were initialized with a uniform distribution having a mean of 0 and a standard deviation of 0.02. 2) The weights corresponding to the RGB channels, excluding pixel-wise positional encoding [3 x 786], were substituted with the parameters from the linear projection layer of the ViTPose-B model. The weights of the volumetric heatmap-based 3D pose estimator were initialized using a normal distribution with a mean of 0 and a standard deviation of 1. Network training was conducted on 4 RTX 3090 GPUs, using the MMPose framework.

Table 1: Comparison with existing methods on the SceneEgo test dataset.

Method	Camera model	Require calibration	MPJPE	PA-MPJPE
Mo^{2}Cap^2 [25]	Scaramuzza *et al.* [16]	yes	92.2	66.01
xR-egopose [20]	N/A	**no**	121.5	98.84
SceneEgo [23]	Scaramuzza *et al.*	yes	89.06	70.10
Wang *et al.* [24]	Scaramuzza *et al.*	yes	**64.19**	**50.06**
Ours	Double Sphere [22]	**no**	68.76	57.73

Table 2: Comparison with existing methods on the our proposed dataset.

Method	Camera model	Require calibration	PA-MPJPE
Wang *et al.*	Scaramuzza *et al.*	yes	156.93
Ours	Double Sphere	**no**	**154.18**

The dataset we propose provides 3D poses from an external camera, which are in a different coordinate system than those typically output by our proposed network. Therefore, training using MPJPE loss is not feasible. Whether in fisheye camera space or external camera space, both use the same units, and applying scale alignment during training could lead to unstable output of 3D poses. Consequently, during fine-tuning, we used a modified PA-MPJPE loss that excludes scaling and only aligns translating and rotating elements through Procrustes alignment.

4.3 Comparison to State-of-the-art Results

To validate the performance of the proposed method, we conducted comparisons with state-of-the-art methods. The methods used for comparison include Mo^{2}Cap^2, xR-egopose, SceneEgo, and Wang *et al.* All four methods were trained using the same dataset to ensure a fair comparison. For methods requiring camera calibration, the dataset-provided calibration data were used. The comparative results are presented in Tab. 1.

Tab. 1 compares the performance of our proposed method with state-of-the-art techniques on the SceneEgo test dataset. The methods used for comparison are indicated based on whether they perform distortion correction using camera calibration and camera models. When comparing the proposed method with the method that does not use a camera model (xR-egopose [20]), the MPJPE decreased from 121.5mm to 68.76mm, and the PA-MPJPE decreased from 98.84mm to 57.73mm, resulting in reductions of 52.73mm and 30.08mm, respectively. Compared to methods that perform camera calibration, the proposed method showed a reduction in MPJPE and PA-MPJPE by 23.44mm and 8.28mm, respectively, for Mo^{2}Cap^2, while it increased by 4.57mm and 7.67 mm, respectively, compared to Wang *et al.*'s method [24]. These results indicate a significant performance improvement for methods that do not perform camera

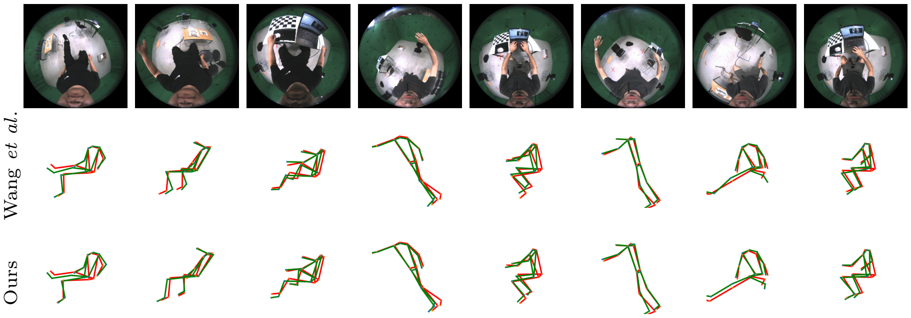


Fig. 4: Qualitative comparison between our method and the previous state-of-the-art method on SceneEgo test dataset. The ground truth pose is shown in red, while the predicted pose is shown in green.

calibration. To compare with calibration-based methods, the proposed method shows improved performance overall, except when compared directly with the method by Wang *et al.* The significant error difference observed with xR-egopose is attributed to the absence of a camera model for error correction and difference in 3D pose estimation techniques. Unlike our method, which predicts 3D poses by jointly estimating depth information from the image, xR-egopose relies solely on 2D keypoints predicted from the image, excluding depth information. Among methods that require camera calibration, the proposed method demonstrates the lowest error, apart from Wang *et al.*'s method. The substantial performance differences between Mo^{2}Cap^2 and SceneEgo can be attributed to differences in network architecture. Both methods use convolutional layer-based backbone networks as feature extractors, but they have different head structures for pose estimation. Mo^{2}Cap^2 and SceneEgo have separate networks for depth prediction, while Wang *et al.*'s method and the proposed method perform simultaneous predictions in a volumetric heatmap format.

Tab. 2 compares the performance of our model with that of Wang et al.'s on our proposed dataset. After fine-tuning, both models were evaluated using PA-MPJPE. Despite not applying calibration, our model outperformed Wang et al.'s. In contrast, Tab. 1 shows Wang et al.'s model performing better in controlled environments, while our model excelled in more realistic settings.

Fig. 4 and Fig. 5 visualize the prediction results of the methods for the SceneEgo test dataset and ours proposed test dataset, respectively, to facilitate a qualitative comparison. These figures visualize the predicted 3D poses and ground truth of the previous state-of-the-art method and our proposed method. In Fig. 4, the predicted poses of the two methods are similar. However, in Fig. 5, the Wang *et al.*'s method [24] shows more accurate prediction results for sitting or crouching poses compared to our method. On the other hand, for poses where the body area is widely spread (such as arms outstretched), our proposed method is relatively more accurate. This occurs because, in poses where the body area is widely spread, the body is often located near the periphery of the image, resulting

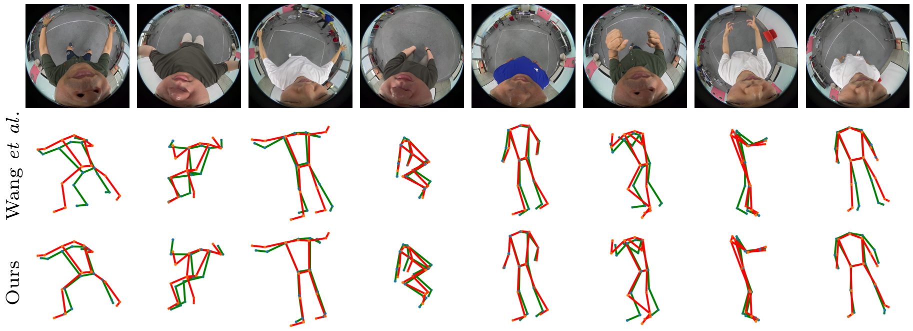


Fig. 5: Qualitative comparison between our method and the previous state-of-the-art method in our proposed test dataset. The ground truth pose is shown in red, while the predicted pose is shown in green.

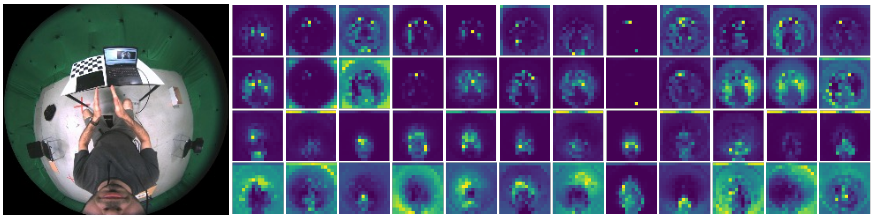
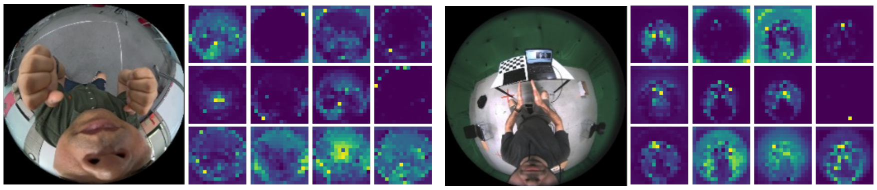


Fig. 6: Visualization results of the self-attention maps from CoordViT. The self-attention maps from the distortion token in CoordViT, visualized to identify focus areas for distortion parameter estimation. Each column represents a different attention head. The first and second rows show the query and key attention maps from the first transformer encoder layer, while the third and fourth rows correspond to the final layer.

in significant radial distortion. Our method can predict distortion parameters for calibration, allowing it to maintain robust performance in poses such as arms outstretched compared to Wang *et al.*'s method.

To evaluate the role of distortion tokens in CoordViT, we utilized the attention map visualization technique introduced in DINO [3]. The visualizations are presented in Fig. 6, depicting the self-attention maps of pose tokens relative to distortion tokens across the layers of CoordViT. The attention maps for queries reveal how each token references others, while the maps for keys show which tokens are referenced by others. Our analysis of these visualizations reveals the following insights: In the first layer of CoordViT (first and second rows), the attention is predominantly directed towards the periphery of the image, focusing on areas relevant to estimating distortion parameters. This suggests an initial emphasis on understanding the fisheye lens effects. In contrast, the attention maps for the final layer (third and fourth rows) show a shift towards regions associated with top masking and pose-related areas of the image. This indicates that as the network progresses through layers, it integrates both pose information and fisheye image masking details for refining distortion parameter estimation.



(a) Our dataset **(b)** SceneEgo test dataset

Fig. 7: Comparison of self-attention maps from the distortion token between different datasets.

Table 3: Evaluation results of the ablation study on SceneEgo test dataset.

Method	MPJPE(/mm)	PA-MPJPE(/mm)
w/o Coord Encoding	72.31	61.44
Scaramuzza's camera model	71.82	60.06
ViT w/ xR-egopose head	141.5	104.78
Ours	**68.76**	**57.73**

In summary, from the visualization of self-attention maps, we observe that in the early layers of the transformer encoder, there is a predominant focus on the outer edges of the image to estimate the distortion token (first row), while in the later layers, the focus shifts to body parts for estimating the distortion token (third row). Additionally, in the later layers, it can be noted that the pose token at the outer edges of the image contributes significantly to the self-attention computation with a focus on the distortion token (fourth row).

To assess whether the proposed method generates consistent attention maps across datasets with different camera parameters, we fine-tuned our model on our dataset and visualized the self-attention maps of the distortion token for both the fine-tuning dataset(Fig. 7a) and the unseen SceneEgo dataset(Fig. 7b). We visualized the first layer of CoordViT, as shown in the second row of Fig. 6. The results indicate that, even without fine-tuning on the SceneEgo dataset, the attention patterns are similar to those from fine-tuning on it, demonstrating the our model's robustness in predicting distortion parameters on unseen data.

4.4 Ablation Study

In this section, we evaluate the effectiveness of the proposed modules in our network by assessing performance changes when specific modules are removed or replaced. We use the SceneEgo test dataset for performance evaluation. Except for the modified modules, all networks are trained and evaluated in the same manner.

Effect of Pixel-wise Positional Encoding in CoordViT. To evaluate the contribution of CoordViT's pixel-wise positional encoding to the egocentric pose estimation pipeline, we assess its impact on performance. In Tab. 3 the item labeld "w/o Coord Encoding" represents the scenario where pixel-wise positional encoding is not applied in CoordViT. The evaluation results indicate a performance decrease when pixel-wise positional encoding is not used, demonstrating its importance for the model's effectiveness.

Effect of Double Sphere Camera Model. We evaluated the impact of different fisheye camera models on performance. In Tab. 3 the item labeled "Scaramuzza's camera model" represents the camera model used in previous methods. Due to the discrepancy in the number of distortion parameters, the distortion parameter estimator was modified accordingly for training. The evaluation results show a decrease in performance when using "Scaramuzza's camera model" indicating its reduced effectiveness compared to the proposed network.

Ablation about Model Architectures. Tab. 1 shows that our model significantly outperforms xR-egopose, another calibration-free method. Since xR-egopose uses a ResNet architecture, we re-implemented it with a ViT backbone for a fair comparison. As shown in Tab. 3 under "ViT w/ xR-egopose head" the ViT version performs worse than the ResNet-based model. This is likely because xR-egopose relies on 2D heatmaps, where ResNet's strength in capturing local features is more effective than ViT's focus on global features.

5 Conclusion

In this paper we propose a novel method for egocentric 3D body pose estimation from single RGB images using the CoordViT network, which mitigates fisheye lens distortion and integrates with a volumetric heatmap-based 3D pose estimator. The network addresses radial distortion by encoding pixel positions, allowing the ViT-based network to extract relevant feature tokens. Distortion parameters are predicted and applied to correct both the input image and the 3D pose. We validated the proposed method on a new dataset, comparing its performance against state-of-the-art models. While it showed slightly lower quantitative performance compared to methods using calibration, it demonstrated superior qualitative performance in handling large motions. We also visualized the self-attention map of the distortion token to highlight the network's focus during distortion correction. Finally, an ablation study confirmed the effectiveness of pixel-wise positional encoding and the Double Sphere camera model for self-calibration in fisheye images.

Acknowledgments. This work was supported by Institute of Information and communications Technology Planning and Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00137, XR User Interaction Evaluation and Employing Its Technology)

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akada, H., Wang, J., Golyanik, V., Theobalt, C.: 3d human pose perception from egocentric stereo videos. arXiv preprint arXiv:2401.00889 (2023)
2. Akada, H., Wang, J., Shimada, S., Takahashi, M., Theobalt, C., Golyanik, V.: Unrealego: A new dataset for robust egocentric 3d human motion capture. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 1–17. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20068-7_1
3. Caron, M., Touvron, H., Misra, I., Jegou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9630–9640 (2021). https://doi.org/10.1109/ICCV48922.2021.00951
4. Habibie, I., Xu, W., Mehta, D., Pons-Moll, G., Theobalt, C.: In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10897–10906 (2019). https://doi.org/10.1109/CVPR.2019.01116
5. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7122–7131. IEEE Computer Society, Los Alamitos, CA, USA (2018). https://doi.org/10.1109/CVPR.2018.00744
6. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5252–5262 (2020). https://doi.org/10.1109/CVPR42600.2020.00530
7. Kolotouros, N., Pavlakos, G., Black, M., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2252–2261 (2019). https://doi.org/10.1109/ICCV.2019.00234
8. Liu, R., Lehman, J., Molino, P., Such, F.P., Frank, E., Sergeev, A., Yosinski, J.: An intriguing failing of convolutional neural networks and the coordconv solution. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9628–9639. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
9. Liu, Y., Yang, J., Gu, X., Chen, Y., Guo, Y., Yang, G.Z.: Egofish3d: Egocentric 3d pose estimation from a fisheye camera via self-supervised learning. IEEE Transactions on Multimedia **25**, 8880–8891 (2023). https://doi.org/10.1109/TMM.2023.3242551
10. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Trans. Graph. **34**(6) (2015). https://doi.org/10.1145/2816795.2818013
11. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5137–5146 (2018). https://doi.org/10.1109/CVPR.2018.00539
12. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2659–2668 (2017). https://doi.org/10.1109/ICCV.2017.288
13. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estdimation from a single rgb image. In: 2019 IEEE/CVF

International Conference on Computer Vision (ICCV). pp. 10132–10141 (2019). https://doi.org/10.1109/ICCV.2019.01023

14. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1263–1272 (2017). https://doi.org/10.1109/CVPR.2017.139
15. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Trans. Graph. **35**(6) (2016). https://doi.org/10.1145/2980179.2980235
16. Scaramuzza, D., Martinelli, A., Siegwart, R.: A toolbox for easily calibrating omnidirectional cameras. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5695–5701 (2006). https://doi.org/10.1109/IROS.2006.282372
17. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2621–2630 (2017). https://doi.org/10.1109/ICCV.2017.284
18. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. pp. 536–553. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_33
19. Tekin, B., Márquez-Neila, P., Salzmann, M., Fua, P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 3961–3970 (2017). https://doi.org/10.1109/ICCV.2017.425
20. Tome, D., Peluse, P., Agapito, L., Badino, H.: xr-egopose: Egocentric 3d human pose from an hmd camera. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7727–7737. IEEE Computer Society, Los Alamitos, CA, USA (2019). https://doi.org/10.1109/ICCV.2019.00782
21. Tome, D., Alldieck, T., Peluse, P., Pons-Moll, G., Agapito, L., Badino, H., de la Torre, F.: Selfpose: 3d egocentric pose estimation from a headset mounted camera. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 6794–6806 (2023). https://doi.org/10.1109/TPAMI.2020.3029700
22. Usenko, V., Demmel, N., Cremers, D.: The double sphere camera model. In: 2018 International Conference on 3D Vision (3DV). pp. 552–560 (2018). https://doi.org/10.1109/3DV.2018.00069
23. Wang, J., Luvizon, D., Xu, W., Liu, L., Sarkar, K., Theobalt, C.: Scene-aware egocentric 3d human pose estimation. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13031–13040. IEEE Computer Society, Los Alamitos, CA, USA (2023). https://doi.org/10.1109/CVPR52729.2023.01252
24. Wang, J., Cao, Z., Luvizon, D., Liu, L., Sarkar, K., Tang, D., Beeler, T., Theobalt, C.: Egocentric whole-body motion capture with fisheyevit and diffusion-based motion refinement. arXiv preprint arXiv:2311.16495 (2023)
25. Xu, W., Chatterjee, A., Zollhöfer, M., Rhodin, H., Fua, P., Seidel, H.P., Theobalt, C.: Mo2cap2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. IEEE Transactions on Visualization and Computer Graphics **25**(5), 2093–2101 (2019). https://doi.org/10.1109/TVCG.2019.2898650
26. Zhang, Y., You, S., Gevers, T.: Automatic calibration of the fisheye camera for egocentric 3d human pose estimation from a single image. In: 2021 IEEE Winter

Conference on Applications of Computer Vision (WACV). pp. 1771–1780. IEEE Computer Society, Los Alamitos, CA, USA (2021). https://doi.org/10.1109/WACV48630.2021.00181

27. Zhou, X., Sun, X., Zhang, W., Liang, S., Wei, Y.: Deep kinematic pose regression. In: Hua, G., Jégou, H. (eds.) Computer Vision – ECCV 2016 Workshops. pp. 186–201. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_17