Odometry-Free Cross-View Pose Fusion for XR Interaction via Visibility-Guided Alignment

Sungjin Hong
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
sjhong0117@etri.re.kr

Cho-rong Yu
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
crryu@etri.re.kr

Youn-Hee Gil
Electronics and Telecommunications Research Institute
Daejeon, Republic of Korea
yhgil@etri.re.kr

Abstract

Egocentric 3D Human Pose Estimation (HPE) suffers from severe perspective distortion and self-occlusion, degrading lower-body pose estimation. In face-to-face XR interactions, a third-person view captured by an interacting partner's front-facing camera provides complementary global context, but the independently moving cameras do not share known cross-view extrinsic parameters or camera odometry. Conventional unweighted Procrustes alignment is sensitive to noisy egocentric joints, which can degrade the estimated rigid transformation. To address this, we propose an odometry-free cross-view fusion framework with Visibility-Guided Alignment (VGA), which incorporates predicted joint visibility as correspondence weights in weighted Procrustes alignment to reduce the influence of unreliable joints during rigid registration. Experiments on a synthetic dataset demonstrate that our method reduces overall MPJPE by 13% compared with the third-person-only baseline.

CCS Concepts

• **Computing methodologies** → **Computer vision problems**.

Keywords

Egocentric 3D pose estimation, cross-view pose fusion, visibility-guided alignment

ACM Reference Format:
Sungjin Hong, Cho-rong Yu, and Youn-Hee Gil. 2026. Odometry-Free Cross-View Pose Fusion for XR Interaction via Visibility-Guided Alignment. In *32nd ACM Symposium on Virtual Reality Software and Technology (VRST '26), November 16–18, 2026, Sendai, Japan.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3822517.3849896

1 Introduction

Accurate 3D Human Pose Estimation (HPE) is important for immersive Extended Reality (XR) interactions. While egocentric cameras enable unconstrained mobility, they introduce severe perspective distortion—enlarging the upper body while foreshortening the lower body—and cause substantial self-occlusion during complex interactions. To address this, Mo2Cap2 [5] utilized cap-mounted fisheye cameras with a lower-body network, while StereoVisPoseNet [1] incorporated pose refinement and visibility estimation into stereo egocentric pose estimation. However, relying on egocentric observations alone remains challenging for accurately recovering full-body pose under severe distortion and self-occlusion.

In face-to-face XR interactions, an interacting partner's front-facing camera can provide complementary third-person skeletal context. Ego+X [3] leveraged visual SLAM and known inter-camera calibration for global pose estimation, while the wearer and partner poses were estimated separately and represented in a common coordinate system for interaction analysis. In our setting, however, independently moving cameras do not share known cross-view extrinsics or camera odometry. Conventional unweighted Procrustes alignment can estimate their relative transformation, but unreliable egocentric joints may bias the alignment. We therefore propose an odometry-free cross-view fusion framework with Visibility-Guided Alignment (VGA), which weights joint correspondences by predicted visibility. We validate our approach on a synthetic dataset, achieving a 13% MPJPE reduction over the third-person-only baseline.

2 Visibility-Guided Cross-View Fusion

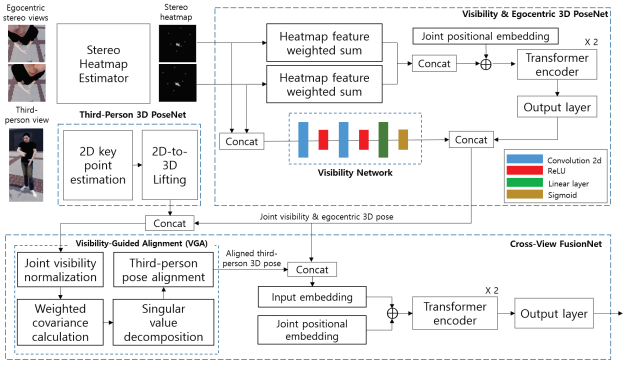


Figure 1: Overall architecture of the proposed Visibility-Guided Cross-View Fusion framework, consisting of stereo heatmap estimation, third-person and egocentric pose estimation, Visibility-Guided Alignment (VGA), and Cross-View FusionNet.

A lightweight Siamese U-Net extracts normalized 2D stereo heatmaps H_L, H_R from self-facing egocentric images. We estimate the third-person 3D pose $P_{third} \in \mathbb{R}^{J\times 3}$ from an interacting partner's front-facing camera using RTMPose [2] and SimpleBaseline [4]. Since P_{third} resides in an independent coordinate system, geometric alignment is required.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
VRST '26, Sendai, Japan
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2811-2/2026/11
https://doi.org/10.1145/3822517.3849896

Table 1: Comparison of Mean Per Joint Position Error (MPJPE, mm) and Procrustes-Aligned MPJPE (PA-MPJPE, mm) across different methods.

Method	MPJPE (Upper)	MPJPE (Lower)	MPJPE ↓ (Full)	PA-MPJPE ↓ (Full)
Ego_{only}	72.52	92.54	82.53	57.66
$Third_{\text{only}}$	65.32	70.55	67.93	46.77
Ego_{refine}	65.35	71.78	68.56	53.11
$Fusion_{\text{w/ Unweighted}}$	59.56	66.60	63.08	48.89
$Fusion_{\text{w/ VGA}}$	54.24	63.49	58.86	44.48

The visibility and egocentric 3D pose module estimates the initial egocentric 3D pose P_{ego} and joint visibility scores V. We employ a vectorized attentional pooling mechanism via spatial softmax to extract joint-specific features from the heatmaps. A Transformer encoder processes these features to predict P_{ego}, while a parallel CNN classifier evaluates the stereo heatmaps to output a visibility confidence score $v_i \in [0, 1]$.

For visibility-guided alignment and cross-view fusion, VGA performs weighted Procrustes alignment over a predefined anchor set $S = \{\text{neck, L/R shoulders, L/R hips}\}$, using predicted visibility v_i as correspondence weights. We compute the weighted cross-covariance matrix $H = \sum_{i \in S} \tilde{v}_i (P_{ego,i} - \mu_{ego})(P_{third,i} - \mu_{third})^T$, where $\tilde{v}_i$ are normalized weights and μ denotes the weighted centroids. SVD of H yields the optimal rotation, while translation is recovered from the weighted centroids. The resulting transformation produces the aligned third-person pose $\hat{P}_{third}$. Finally, the Cross-View FusionNet processes the concatenated 7-dimensional tokens $F_i = [P_{ego,i} \parallel \hat{P}_{third,i} \parallel v_i]$ using a Transformer encoder to output the refined 3D pose P_{final}.

3 Experiment

We constructed a synthetic dataset of synchronized stereo egocentric and third-person RGB images with 3D coordinates and visibility labels for 16 joints (head, neck, shoulders, elbows, wrists, hips, knees, ankles, and toe tips). The dataset includes 10 characters, 8 backgrounds, and diverse lighting conditions, with 18,535 training and 4,634 evaluation samples.

For training, we use Adam with MSE loss ($lr = 10^{-4}$) to train the Stereo Heatmap Estimator, Egocentric 3D PoseNet, and Cross-View FusionNet for 200, 300, and 100 epochs, respectively. Joint VisibilityNet is trained for 200 epochs with BCE loss, while the 2D-to-3D lifting model [4] is trained for 500 epochs ($lr = 10^{-3}$) using pretrained RTMPose [2].

The limitations of egocentric-only refinement are evident in Table 1. Ego_{only} exhibits substantially higher lower-body than upper-body error (92.54 vs. 72.52 mm), reflecting severe perspective distortion and self-occlusion. Ego_{refine}, which refines P_{ego} using only the egocentric pose and visibility scores without third-person input, reduces MPJPE to 68.56 mm. However, its PA-MPJPE remains higher than $Third_{\text{only}}$ (53.11 vs. 46.77 mm), indicating that egocentric-only refinement is insufficient to fully recover global pose structure.

Incorporating the aligned third-person pose via conventional unweighted Procrustes alignment ($Fusion_{\text{w/ Unweighted}}$) further reduces Full MPJPE to 63.08 mm, outperforming Ego_{only}, Ego_{refine}, and $Third_{\text{only}}$. This demonstrates that the improvement is not solely due to additional refinement, but arises from the complementary information provided by the third-person view.

Despite improving MPJPE, unweighted Procrustes alignment remains sensitive to unreliable egocentric joint correspondences, resulting in a PA-MPJPE of 48.89 mm, which is higher than that of $Third_{\text{only}}$ (46.77 mm). This suggests that naive cross-view fusion improves absolute pose accuracy but does not necessarily improve pose accuracy after Procrustes alignment. Our proposed VGA addresses this limitation by weighting joint correspondences according to their predicted visibility during alignment. Joint VisibilityNet achieves 89.12% visibility classification accuracy. Consequently, $Fusion_{\text{w/ VGA}}$ achieves the best PA-MPJPE of 44.48 mm and Full MPJPE of 58.86 mm, corresponding to a 13% reduction compared with $Third_{\text{only}}$. While the evaluation is limited to synthetic data, the results demonstrate the benefit of visibility-guided alignment under controlled cross-view conditions.

4 Conclusion and Future Work

We presented an odometry-free cross-view fusion framework with Visibility-Guided Alignment (VGA), which uses predicted joint visibility in weighted Procrustes alignment to robustly align egocentric and third-person poses without known cross-view extrinsics. VGA reduces MPJPE by 13% over the third-person-only baseline. Future work will focus on bridging the synthetic-to-real domain gap through domain adaptation and validating the framework in real-world XR interactions.

Acknowledgments

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [26ZC1100, Development of Spatial Media Technology and Interaction Technology for Convergence of the Real and Virtual World].

References

[1] Sungjin Hong, Hye-sun Kim, Cho-rong Yu, and Youn-Hee Gil. 2025. StereoVisPoseNet: Stereo-based Visibility-aware Egocentric 3D Pose Estimation Network. In *Proceedings of the 2025 31st ACM Symposium on Virtual Reality Software and Technology*. 1–2.

[2] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. 2023. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399* (2023).

[3] Yuxuan Liu, Jianxin Yang, Xiao Gu, Yao Guo, and Guang-Zhong Yang. 2022. Ego+X: An Egocentric Vision System for Global 3D Human Pose Estimation and Social Interaction Characterization. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 5271–5277. doi:10.1109/IROS47612.2022.9981710

[4] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *2017 IEEE international conference on computer vision (ICCV)*. IEEE, 2659–2668.

[5] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Pascal Fua, Hans-Peter Seidel, and Christian Theobalt. 2019. Mo 2 cap 2: Real-time mobile 3d motion capture with a cap-mounted fisheye camera. *IEEE transactions on visualization and computer graphics* 25, 5 (2019), 2093–2101.