

Supplementary Material for HOIGaze: Gaze Estimation During Hand-Object Interactions in Extended Reality Exploiting Eye-Hand-Head Coordination

ZHIMING HU*, University of Stuttgart, Germany The Hong Kong University of Science and Technology (GZ), China
 DANIEL HAEUFLE, University of Tuebingen, Germany The Center for Bionic Intelligence Tuebingen Stuttgart, Germany
 SYN SCHMITT, University of Stuttgart, Germany The Center for Bionic Intelligence Tuebingen Stuttgart, Germany
 ANDREAS BULLING, University of Stuttgart, Germany

ACM Reference Format:

Zhiming Hu, Daniel Haeufle, Syn Schmitt, and Andreas Bulling. 2025. Supplementary Material for HOIGaze: Gaze Estimation During Hand-Object Interactions in Extended Reality Exploiting Eye-Hand-Head Coordination. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '25)*, August 10–14, 2025, Vancouver, BC, Canada. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3721238.3730692>

1 ATTENDED HAND RECOGNITION ACCURACY AND GAZE ESTIMATION PERFORMANCE

Our attended hand recogniser achieves an accuracy of 83.0% on HOT3D (Cross-User), 83.6% on HOT3D (Cross-Scene), and 86.5% on ADT, respectively. To better understand the error distribution of our method, we further calculated the mean angular errors of different methods when the recognised attended hand is correct or wrong. We can see from Table 1 that our method significantly outperforms other methods when the recognised attended hand is correct, achieving an improvement of 17.5% (9.03° vs. 10.94°) on HOT3D (Cross-User), 14.3% (8.27° vs. 9.65°) on HOT3D (Cross-Scene), and 8.1% (8.52° vs. 9.27°) on ADT. We also find that when the recognised attended hand is wrong, our method can achieve superior or comparable performance with the state of the art, demonstrating the robustness of our method.

2 ABLATION STUDY

Cross-Modal Transformer. We removed the self-attention and cross-attention used in the cross-modal Transformers respectively to re-train our method. The results in Table 2 show that our method

*Corresponding author

Authors' addresses: Zhiming Hu, University of Stuttgart, Germany and The Hong Kong University of Science and Technology (GZ), China; Daniel Haeufle, University of Tuebingen, Germany and The Center for Bionic Intelligence Tuebingen Stuttgart, Germany; Syn Schmitt, University of Stuttgart, Germany and The Center for Bionic Intelligence Tuebingen Stuttgart, Germany; Andreas Bulling, University of Stuttgart, Germany.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGGRAPH Conference Papers '25, August 10–14, 2025, Vancouver, BC, Canada

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1540-2/2025/08

<https://doi.org/10.1145/3721238.3730692>

Table 1. Attended hand recognition accuracies and mean angular errors of different methods on the HOT3D and ADT datasets when the recognised attended hand is correct (✓) or wrong (×). Best results are in bold.

	HOT3D-User	HOT3D-Scene	ADT
Recognition Accuracy	83.0%	83.6%	86.5%
<i>Head Direction</i> ✓	22.96°	23.11°	22.25°
<i>DGaze</i> ✓	14.27°	12.79°	9.88°
<i>FixationNet</i> ✓	13.95°	12.52°	9.91°
<i>Pose2Gaze</i> ✓	10.94°	9.65°	9.27°
Ours ✓	9.03°	8.27°	8.52°
<i>Head Direction</i> ×	24.38°	23.66°	22.28°
<i>DGaze</i> ×	14.40°	12.87°	10.21°
<i>FixationNet</i> ×	14.25°	12.55°	9.99°
<i>Pose2Gaze</i> ×	11.89°	10.54°	9.80°
Ours ×	11.01°	10.53°	10.47°

Table 2. Mean angular errors of our method's different ablated versions on the HOT3D and ADT datasets. Best results are in bold.

	HOT3D-User	HOT3D-Scene	ADT
w/o self-attention	9.80°	8.68°	9.08°
w/o cross-attention	10.22°	8.98°	8.99°
Ours	9.37°	8.64°	8.78°
recogniser GCN 0	9.75°	8.78°	8.99°
recogniser GCN 1	9.56°	8.74°	8.89°
recogniser GCN 2 (Ours)	9.37°	8.64°	8.78°
recogniser GCN 4	9.60°	8.86°	8.96°
recogniser GCN 8	9.69°	8.71°	8.96°
estimator GCN 0	9.87°	9.13°	9.20°
estimator GCN 1	10.24°	9.01°	9.11°
estimator GCN 2	9.85°	8.66°	8.74°
estimator GCN 4 (Ours)	9.37°	8.64°	8.78°
estimator GCN 8	9.69°	8.65°	8.84°
scene object 0	10.34°	9.19°	9.03°
scene object 1 (Ours)	9.37°	8.64°	8.78°
scene object 2	9.59°	8.61°	8.91°
scene object 3	9.88°	8.72°	9.36°
scene object 4	9.79°	8.68°	9.16°

achieves significantly better performances than the ablated versions (paired Wilcoxon signed-rank test, $p < 0.01$), demonstrating

that both the self-attention and cross-attention help improve our method's performance.

GCN in Attended Hand Recogniser and Gaze Estimator. We changed the number of residual GCN layers used in our attended hand recogniser and gaze estimator respectively to re-train our method. We can see from the results in Table 2 that using two residual GCNs in the attended hand recogniser and four residual GCNs in the gaze estimator achieves the best performance. One exception is that using two residual GCNs in the gaze estimator achieves better performance than using four residual GCNs on the ADT dataset (8.74° vs. 8.78°).

This is because we trained our method on ADT using a static hand gesture that requires fewer GCN layers to process.

Scene Object Number. We changed the number of scene objects to re-train our method. As can be seen from the results in Table 2 that using the nearest scene object achieves the best performance. One exception is that using the nearest two objects achieves better performance than using one object on the HOT3D dataset for cross-scene evaluation (8.61° vs. 8.64°). This is because different environments usually have different scene layouts and thus our method needs more scene object information to improve its generalisation ability for different environments.