GazeD: Context-Aware Diffusion for Accurate 3D Gaze Estimation

Supplementary Material

1. Additional qualitative results

We report qualitative results for the three datasets (GFIE [4], GAFA [6] and Ego-Gaze) in different scenarios, with different camera angles and subject distances from the camera. In Figure 1 we demonstrate some scenarios in which the ”Context with objects” helps the model to get a more accurate result. In Figure 2, 3, 4, the first column shows the original image, the second one shows the predicted and GT gaze direction, and the third compares the GT pose with the predicted one. Every prediction is the average of the 20 hypotheses output of the diffusion model. In Figure 5 a visual comparison between GazeD and Gaze360 [5] and Hu *et al.*[4] is reported.

2. Dataset preprocessing

2.1. GFIE

GFIE dataset does not provide 2D/3D human poses. Every dataset sample is composed of a pair of rectified RGB and depth images and gaze annotations. Since the depth maps are noisy with lots of missing values, computing only the 2D pose and then using depth values to obtain the 3D pose was not a good solution. We then decided to use metrabs [7] to estimate both the 2D and 3D poses.

2.2. GAFA

GAFA dataset [6] is provided with only 2D poses computed with an old version of OpenPose [1]. As GAFA dataset is a multiview dataset, the optimal solution to get 3D poses was to triangulate multi-view keypoints using intrinsics and extrinsics provided by the dataset, applying a RANSAC triangulation method [2] and then transforming the resulting pose from the world reference frame to each camera frame. Despite the presence of these poses, GAFA dataset contains several complex scenarios, with the subjects often occluded; openpose annotations often included missing keypoints, resulting in inaccurate triangulated human poses. For this reason, 2D poses were recomputed with metrabs, which can better handle these cases. If some 3D poses were not computable due to limited views, those poses were discarded from the training set, while in the test set, body and head locations present in the annotations of the dataset were used to substitute the fundamental joints to run our method (pelvis and eye midpoint) and gaze joint were added to ensure that all test samples were included.

2.3. Ego-Gaze

Ego-Gaze dataset is realized starting from Ego-Exo [3] dataset, which includes Aria Glasses[1] eye-tracking data. However, despite the presence of these annotations, the dataset does not include a dedicated subset specifically designed for the gaze estimation task. Therefore, we created a subset starting by collecting exocentric frames from the Ego-Pose subset, which comprises automatic and manual annotations of 2D Poses and 3D triangulated poses. We picked frames with manual annotations to ensure the best possible poses. Since manual annotations have missing keypoints, only poses with all the joints were included in our subset. Gaze was inserted as a joint in all the poses just by converting the rotation angles reported in the aria files into normalized Cartesian coordinates and rotating them in the camera reference frame using the camera poses. Since the pose annotation is not provided with a center pelvis, it was manually added as the mean point between the left and the right hip, just to have the root joint. To enable a comparison with competitors, also head crops were needed. Since the dataset doesn’t provide them, and using face joints to crop the images was not an efficient solution in many cases, we trained a YOLOv8[2] with the Hollywood Head dataset [8] to detect heads in the scene. The Yolov8 was run in inference with all the videos of Ego-Pose to ensure that sequential frames were produced for a temporal method such as Gaze360. After obtaining the head crops, only samples with the available complete 3D pose were kept. The final Ego-Gaze subset comprises about 157k frames for the training set, 20k for validation and 45k for testing.

[1] www.projectaria.com

[2] https://github.com/ultralytics/ultralytics

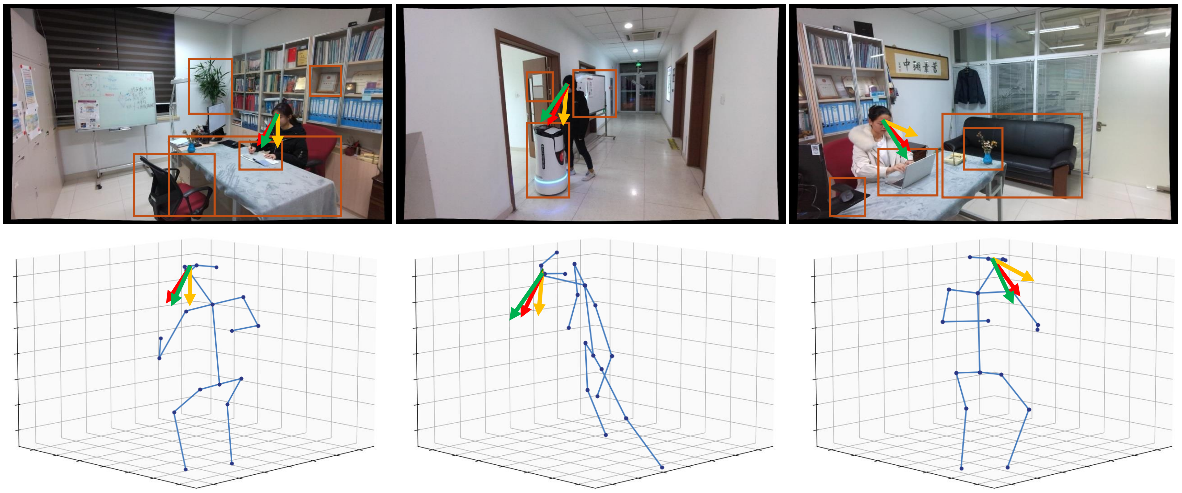

Figure 1. Qualitative comparison of GazeD with and without the *Context with objects* module. Ground truth data is shown in green, predictions from GazeD in red, and results from GazeD without *Context with objects* in yellow.

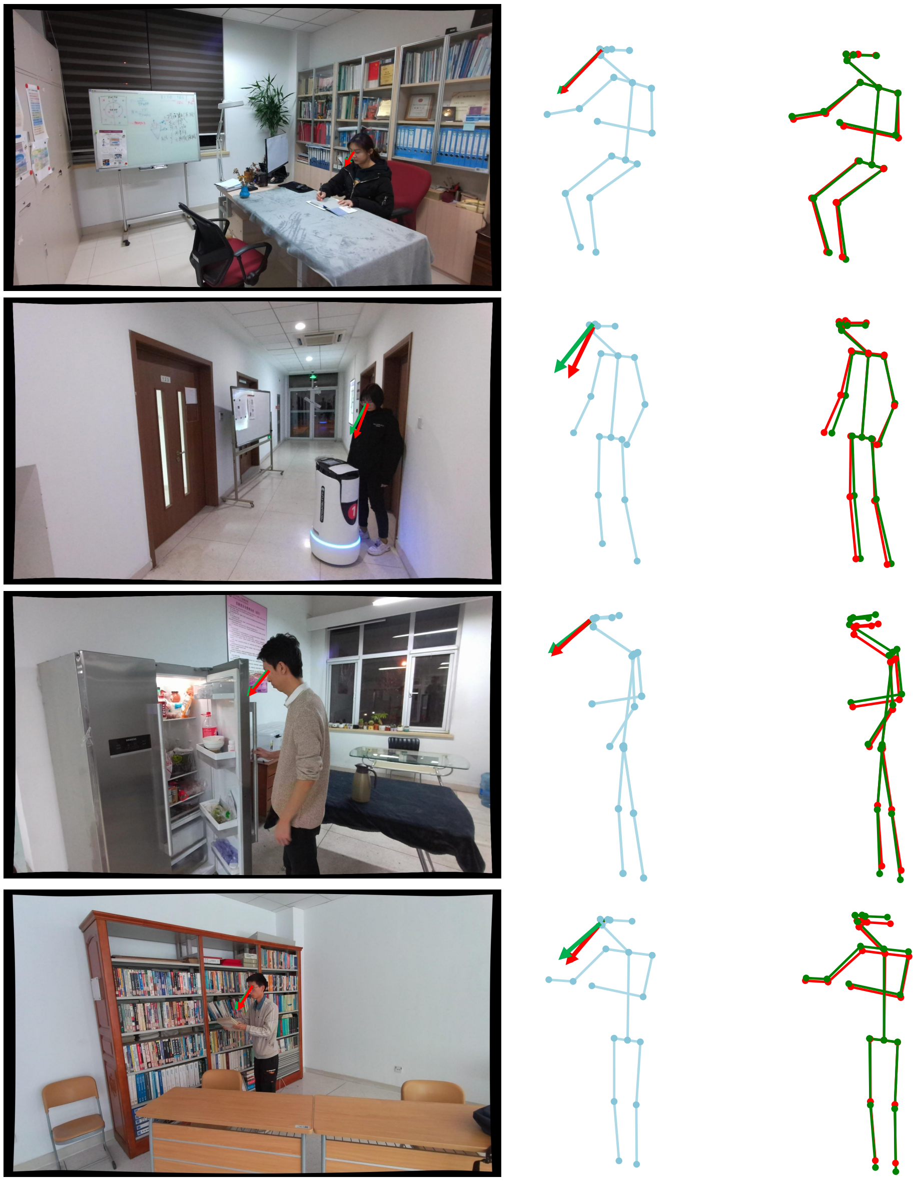

Figure 2. Qualitative results of GazeD on the GFIE dataset. Ground truth data is shown in green, predictions from GazeD in red.

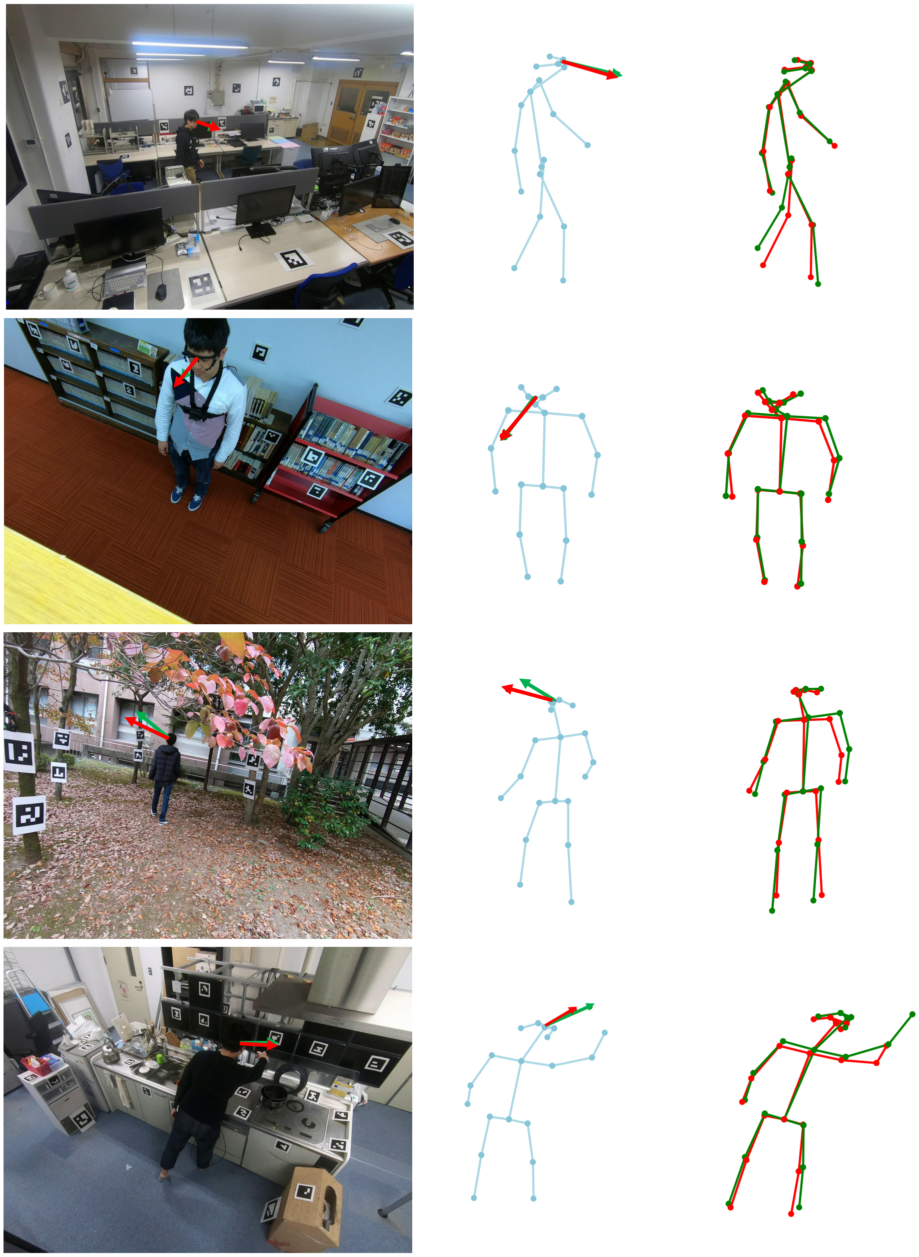

Figure 3. Qualitative results of GazeD on the GAFA dataset. Ground truth data is shown in green, predictions from GazeD in red.

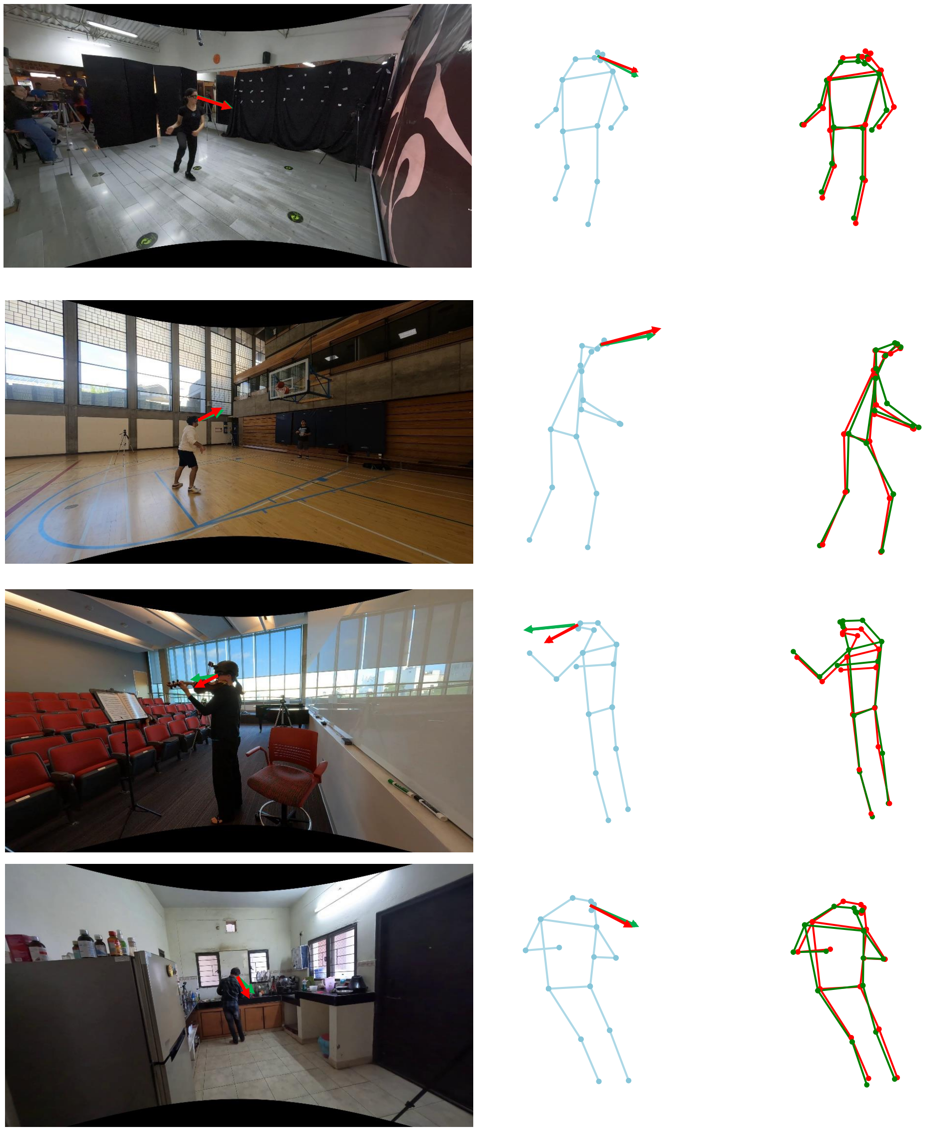

Figure 4. Qualitative results of GazeD on the Ego-Gaze dataset. Ground truth data is shown in green, predictions from GazeD in red.

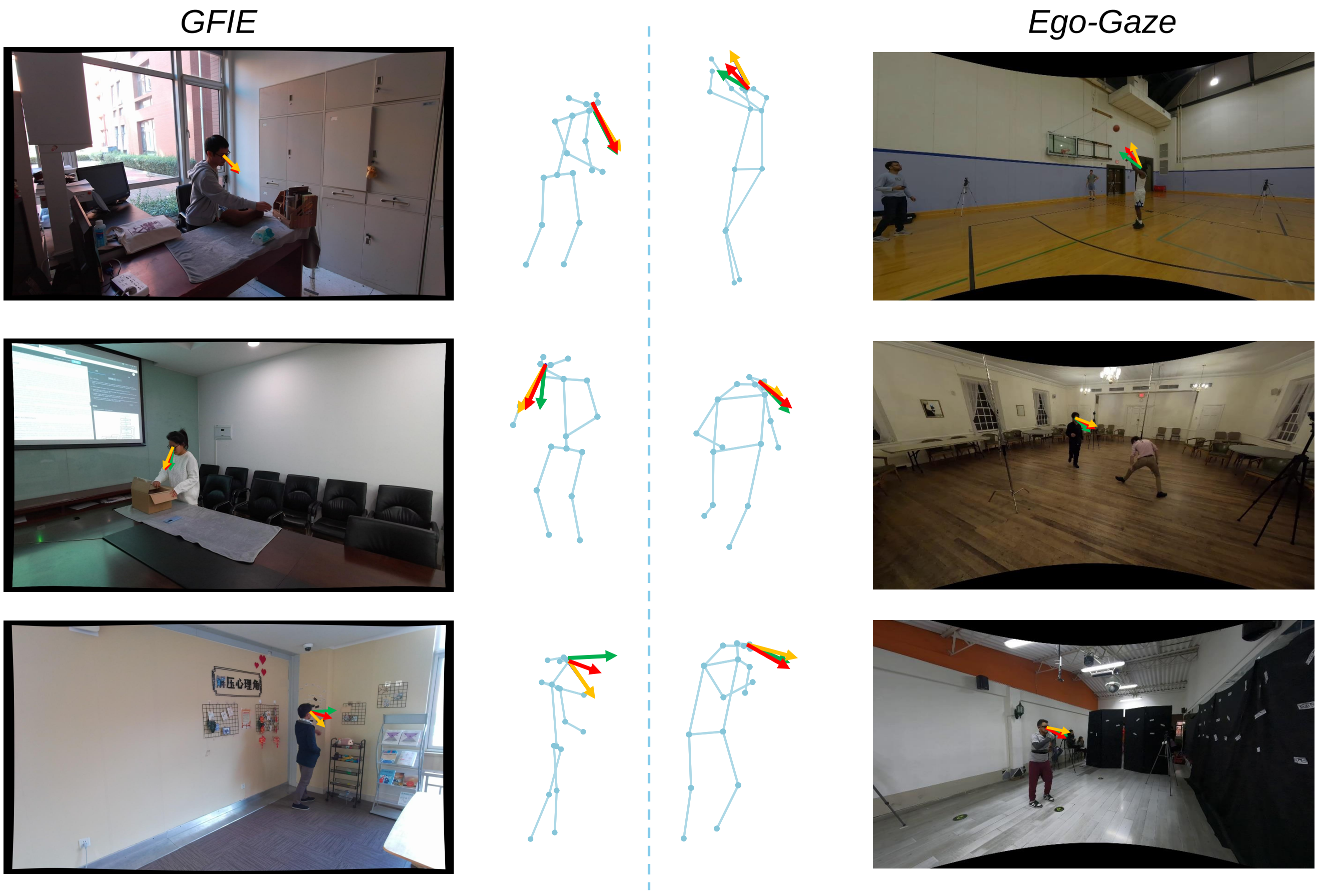

Figure 5. Qualitative comparison with competitors. Ground truth data is shown in green, predictions from GazeD in red, and in yellow are reported the predictions of Gaze360 [5] (right) and [4] (left).

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[2] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, 2018.

[3] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

[4] Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *CVPR*, 2023.

[5] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019.

[6] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *CVPR*, 2022.

[7] István Sárándi, Alexander Hermans, and Bastian Leibe. Learning 3D human pose estimation from dozens of datasets using a geometry-aware autoencoder to bridge between skeleton formats. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023.

[8] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev. Context-aware cnns for person head detection. In *ICCV*, 2015.