Weakly-Supervised Physically Unconstrained Gaze Estimation: Supplementary Material

Rakshit Kothari[1,2*] Shalini De Mello[1] Umar Iqbal[1]
Wonmin Byeon[1] Seonwook Park[3] Jan Kautz[1]
[1]NVIDIA [2]Rochester Institute of Technology [3]Lunit Inc.
rsk3900@rit.edu; spark@lunit.io
{shalinig, uiqbal, wbyeon, jkautz}@nvidia.com

Overview

In this supplementary document, we show additional experimental results and provide more implementation details. Specifically, we demonstrate the advantage of using weak labels from LAEO data on an additional in-the-wild physically unconstrained gaze-related task besides gaze estimation. For this we incorporate our gaze estimation pipeline from AVA-LAEO into the current state-of-the-art visual target estimation network [1] (termed "VATnet" here) and evaluate its performance. Next, for the task of physically unconstrained gaze estimation, we provide additional ablation experiments (besides those in Sec. 4.1 of the main paper), including for the aleatoric and symmetry losses; for various formulations of the pseudo gaze and geometric 3D LAEO losses; and for the utility of the geometric 2D LAEO loss. We show more performance details of the various training datasets used in the cross-dataset experiments (in Sec. 4.2 of the main paper) for different gaze yaw angles. Finally, we provide more details of pre-processing the CMU Panoptic and AVA-LAEO datasets, and analyze the reliability of the 3D gaze labels extracted from real-world LAEO data.

A. Weakly-Supervised Visual Target Estimation

Chong *et al*. [1] proposed a novel spatio-temporal architecture (VATnet), which predicts fixation targets of subjects within a given video frame. In this experiment, we explore if LAEO-based weakly-supervised 3D gaze helps to estimate more accurate visual targets as well. We use LAEO 3D gaze estimation as an auxiliary task while training networks for visual target estimation in a semi-supervised setting. This provides additional weak gaze annotations from the noisy, in-the-wild AVA-LAEO dataset.

*Rakshit Kothari was an intern at NVIDIA during the project.

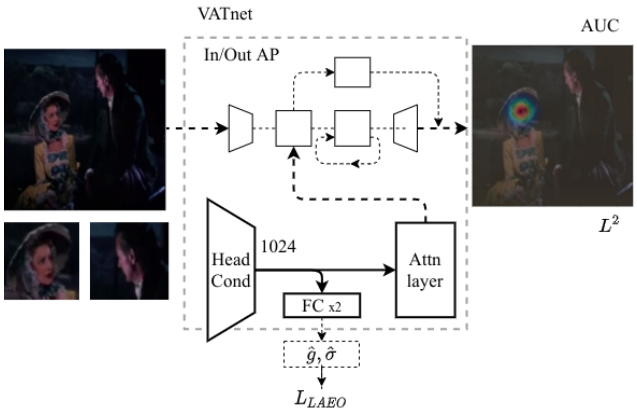

Figure 1: A simple modification of the VATnet architecture [1]. Two fully connected layers serve as an auxiliary task to predict 3D gaze from the head conditioning branch of the original VATnet architecture. The LAEO losses (see Section 3.3 in the main paper) on the predicted gaze vectors for the AVA-LAEO dataset are then used to fine-tune the final layer of the head conditioning branch. Facial features extracted from the fine-tuned head conditioning branch then proceed to VATnet for the visual attention target prediction task. Please refer to Chong *et al*. [1] for a full description of their network architecture.

Method VATnet comprises of four modules, a head conditioning branch, which generates gaze-related features from an input head image; a main scene branch, which generates scene-related feature maps based on the saliency of an input scene image; a recurrent attention prediction module, which fuses gaze- and scene-related features across contiguous video frames; and lastly, a heatmap conditioning branch, which generates a visual target prediction heatmap (see Fig. 1). VATnet's head conditioning branch is a ResNet-50 module initialized with weights from a gaze estimation network trained on the EYEDIAP dataset [2].

	AUC (↑)	$L2$ Dist (↓)	*out-of-frame* AP (↑)
VAT	0.846	0.141	**0.861**
VAT + AVA-LAEO	**0.865**	**0.136**	0.855
Human	0.921	0.051	0.925

Table 1: Improvements to the VATnet baseline [1] by adding weak supervision from the AVA-LAEO dataset using the best configuration of LAEO loss functions described in Table 1 of the main paper.

Utilizing this gaze estimator, Chong *et al*. [1] demonstrate state-of-the-art results on a new dataset called *VisualAttentionTarget*, which comprises of annotated gaze target locations on the image plane. In our experiments we jointly train this VATNet architecture with both the training set of the original fully-supervised VAT dataset and with the AVA-LAEO dataset. To do so, we modify the VATNet architecture and add two fully connected layers to the output of the head conditioning branch, and train it to additionally predict weak 3D gaze vectors derived from the AVA-LAEO dataset (see Fig. 1). We train with samples from AVA-LAEO using the LAEO loss $L_{SYM} + L^{2D}_{geom} + L^{3D}_{geom} + L^{pseudo}_{G}$ only.

Data Preparation VATnet requires three input modalities. First, it requires a full scene image with known head bounding box locations for each annotated subject. Next, it requires a 2D pixel gaze target location on the image plane for the said subject and finally, an *in-out* label, which indicates if the target is within or out of a frame. For this task, to use the LAEO data we input the same 7-frame sequence centered around a LAEO annotation. We treat the 2D cyclopean eye P^{2D} (see the sub-section titled "Scene Geometry Estimation" within Sec. 3.3 of the main paper) of subject B as the target for subject A and vice versa for subject B. The nature of the AVA-LAEO data ensures that all target locations are within an image frame and we assume this to be the default *in-out* ground truth state. We do not pre-process or augment the AVA-LAEO data and directly re-train Chong *et al*.'s original implementation of VATnet with the two datasets with minimal modifications.

Results Following Chong *et al*. [1], we evaluate the area under the curve (AUC) for correct target location prediction (within a pre-specified distance threshold on the image plane), the $L2$ distance between the predicted and ground truth target locations in the scene and the out-of-frame prediction's average precision (AP). We report the scores on the VAT test dataset, averaged across training epochs 2-30, both for the author's original method [1] and our proposed modification. Table 1 shows the benefits of jointly training with the AVA-LAEO and VAT datasets. We notice an improvement in the AUC and $L2$ distance metrics for visual target prediction. These encouraging results suggest that weak supervision from noisily-labeled in-the-wild LAEO data can potentially also aid other gaze-related tasks, *e.g.*, visual attention target prediction besides 3D gaze estimation. We also note a reduction in the *out-of-frame* AP, which is not surprising as all target locations for a given subject in the AVA-LAEO dataset lie within image bounds and hence it provides labels for only one (*i.e.*, the in-frame) class.

	Temporal		Static	
	Frontal face crops °	All head crops °	Frontal face crops °	All head crops °
Pinball	10.38	13.77	11.4	15.62
Aleatoric	9.8	13.65	11.14	15.24
Pinball$_{+L_{sym}}$	10.05	13.37	11.04	15.35
Aleatoric$_{+L_{sym}}$	**9.79**	**12.94**	**10.94**	**15.07**

Table 2: Summary of performance gain by employing an aleatoric gaze loss (described in Sec. 3.3, "Aleatoric Gaze Loss" of the main paper) and the effects of incorporating a symmetry constraint (described in Sec. 3.3, "Symmetry Loss" of the main paper). All values reported are angular gaze errors in degrees (lower is better) for the fully-supervised within-dataset experiment on Gaze360.

B. Additional Ablation Studies

For the task of physically unconstrained gaze estimation, we provide additional ablation experiments besides those in Sec. 4.1 of the main paper.

B.1. Aleatoric and Symmetry Losses

In the normalized eye co-ordinate system [5], where the z axis passes through the 3D cyclopean eye center of each face, constraining gaze yaw prediction to be equal and opposite for a face and its symmetrically flipped version, is an intuitive constraint, which can be employed during training. Our experiments show that using this symmetry constraint and the aleatoric gaze loss improve the baseline performance of [5] on both variants of the author's original fully-supervised ResNet-18-based gaze estimator (temporal and static), which use the pinball gaze loss. Table 2 shows a detailed comparison of the effects of adding the symmetry constraint to the pinball (from [5]) and aleatoric (ours) loss functions for a within-dataset fully-supervised experiment on Gaze360. Here we train our gaze network with Gaze360's entire training set (with its gaze labels) and evaluate it on Gaze360's test set. Note that the symmetry constraint improves the performance of both the pinball and aleatoric losses.

We also observe that for this within-dataset experiment, the aleatoric loss consistently outperforms the pinball loss and that the combination of the aleatoric and symmetry losses results in the best overall performance (Table 2). In addition to this, we observe that the aleatoric loss also outperforms the pinball loss in the cross-domain purely weakly-supervised experimental setting. By replacing the

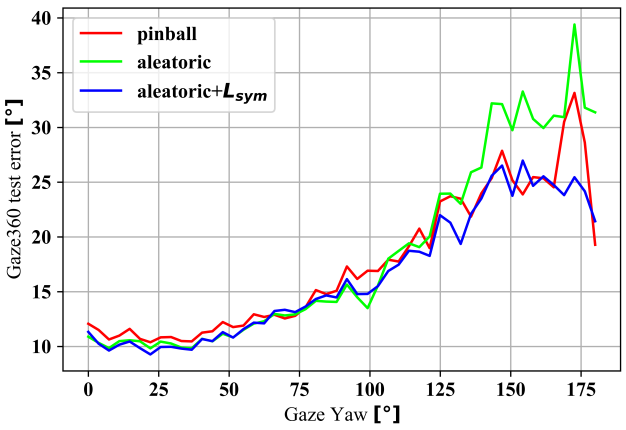

Figure 2: Gaze360 test error (in degrees) as a function of gaze yaw for the fully-supervised within-dataset experimental on Gaze360. Note that gaze error increases as faces turn away from the camera.

aleatoric loss with the pinball loss (from [5]), our best temporal network (trained with all the LAEO losses and corresponding to the last row of Table 1 in the main paper), generalizes less effectively to Gaze360. For AVA-LAEO its gaze error of 26.3° increases to 28.7° and for CMU Panoptic it increases from 25.9° to 26.1°.

B.2. Variants of L_G^{pseudo}

The LAEO activity provides us with the constraint that the predicted 3D gaze from subjects A and B in LAEO must be equal and opposite in a shared camera coordinate system. There are multiple ways in which we can implement this constraint. As an ablation, we explore two additional formulations for this LAEO constraint besides the one described in Sec. 3.3 titled "Pseudo Gaze LAEO Loss" of the main paper: a) naive LAEO enforcement and b) using the most confident gaze prediction for a pair of faces in LAEO as the pseudo ground truth gaze direction. In either experiment, we replace the L_G^{pseudo} loss in our best (temporal) purely weakly-supervised cross-dataset configuration that is trained with all the LAEO losses $L_{sym}+L_G^{pseudo}+L_{geom}^{2D}+L_{geom}^{3D}$ (corresponding to the last row in Table 1 of the main paper) with one of these losses.

Naive LAEO Enforcement Here we naively enforce the predicted vectors $\hat{g}_A^{3D}$ and $\hat{g}_B^{3D}$ to be equal and opposite by minimizing the resultant angular cosine distance between $\hat{g}_A^{3D}$ and $-\hat{g}_B^{3D}$. In this constraint, predictions for both faces could be modified by the network. In order to achieve this, our gaze estimation network could either improve its prediction for the difficult face in a LAEO pair (see Fig. 2, which show that gaze prediction error increased with extreme gaze angles), or it could deteriorate its prediction for the clearer frontal face to satisfy this naive LAEO objective. Our experiments show a reduction in cross-dataset performance on the entire Gaze360 test set (CMU Panoptic: 25.9° → 28.2° and AVA-LAEO: 26.3° → 26.9°) with this naive variant of the LAEO loss versus the one described in Sec. 3.3 of the main paper.

Confident Gaze Prediction In this experiment, we regard the more confident of the two predicted gaze vectors for a LAEO pair as the pseudo ground truth g_{pseudo}^{3D} gaze label as opposed to their weighted average used in Sec. 3.3 of the main paper. That is, $g_{pseudo}^{3D} = \hat{g}_A^{3D}$ if $W_A \geq W_B$ (from Eq. 1 in the main paper) and vice versa for subject B. Our experiments show a reduction in cross-dataset performance with this variant of the LAEO pseudo ground truth label as well versus the one used in Sec. 3.3 of the main paper (CMU Panoptic: 25.9° → 27.24° and AVA-LAEO: 26.3° → 27.8°).

B.3. Variant of L_{geom}^{3D}

We also compare the performance of our L_{geom}^{3D} loss formulation used in Sec. 3.3 of the main paper to a conventional 3D angular cosine loss, whose ground truth is assumed to be along the line joining LAEO subjects' estimated 3D eyes. Empirically, we observe that replacing L_{geom}^{3D} with a cosine loss in our best (temporal) purely weakly-supervised configuration (last row of Table 1 in the main paper), results in consistently worse performance on Gaze360 (CMU Panoptic: 25.9° → 30.0° and AVA-LAEO: 26.3° → 29.63°).

B.4. Utility of L_{geom}^{2D}

The 2D eye position on the image plane can be estimated without depth ambiguity and is more reliable than the 3D eye position. To quantify the contribution of L_{geom}^{2D} to the overall performance of our system, we add increasing noise (z-only) as a ratio of the absolute ground truth depth of the 3D eye positions to subjects under LAEO in the CMU Panoptic dataset, train various purely weakly-supervised configurations (as described in Sec. 4.1 of the main paper) with and without L_{geom}^{2D} and evaluate on Gaze360 (Fig. 3). While we see gaze prediction accuracy deteriorate with increasing depth noise, the inclusion of L_{geom}^{2D} constrains gaze ambiguity and reduces the degradation of gaze estimates. Besides this, we also observe that including L_{geom}^{2D} makes gaze predictions more consistent and reduces the standard deviation of errors on Gaze360's test set (CMU Panoptic: 27.0° → 23.7° and AVA-LAEO: 23.6° → 19.8°).

C. Detailed Cross-dataset Performance

For the cross-dataset experiment described in Sec. 4.2 and Table 2 of the main paper, we additionally analyze the variation in gaze errors with varying gaze yaw angles on the Gaze360 test set. We consider the case of training with (a) GazeCapture only (dashed curves in Fig. 4) or (b) with GazeCapture and AVA-LAEO in (solid curves Fig. 4). The corresponding curves for training with (a) ETH-XGaze

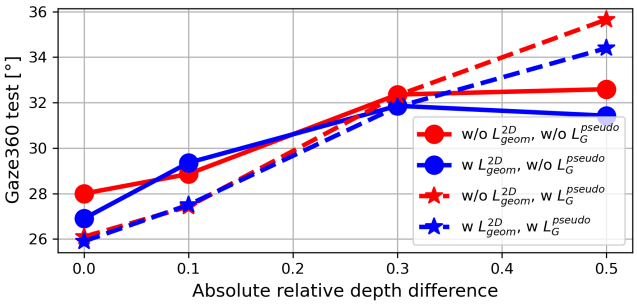

Figure 3: Purely weakly-supervised performance of CMU Panoptic on Gaze360, with added relative depth noise ($\mu = 0$, $\sigma = \{0.1, 0.3, 0.5\}$), when trained with different combinations of LAEO losses (L_{geom}^{3D} is always on). With the L_{geom}^{2D} loss included, performance degrades more gracefully on increasing depth noise versus without. Plots show median values across 4 different training runs initialized with different network weights.

only or (b) with ETH-XGaze and AVA-LAEO are shown in Fig. 5. The blue curves show performance on the entire Gaze360 test set, while the red curves are for its subset containing frontal faces only.

The AVA-LAEO dataset exhibits a large distribution of extreme gaze angles as the LAEO activity largely consists of people with side profiles fixating at each other (see Fig. 1 and Fig. 2 in main paper and Fig. 7 in the supplementary for examples). This conveniently augments datasets with narrow gaze distributions, *e.g.*, GazeCapture (dashed versus solid curves in Fig. 4), which is largely concentrated about gaze pitch and yaw values of zero (from Fig. 3 of the main paper) and helps them generalize better to Gaze360. The AVA-LAEO dataset also contains a large appearance variability because of being collected from in-the-wild videos, which positively augments datasets collected indoors only, *e.g.*, ETH-XGaze (dashed versus solid curves Fig. 5) and helps it generalize better to Gaze360 as well. On jointly training either the GazeCapture or ETH-XGaze dataset with AVA-LAEO, we see a significant boost in their performance on all head crops from Gaze360, including faces with large profile views (blue curves in Fig. 4 and Fig. 5). Interestingly, adding the AVA-LAEO dataset improves cross-domain performance of GazeCapture and ETH-XGaze on Gaze360's frontal face crops as well (red curves in Fig. 4 and Fig. 5).

D. Data Pre-processing

We first describe in detail how we pre-process the CMU Panoptic (*haggling* activity subset) and the AVA-LAEO datasets. Then we analyze the effect of the simplifying assumptions that we employed to estimate scene geometry (as described in Sec. 3.3 of the main paper) on the reliability of 3D gaze annotations derived from real-world LAEO data.

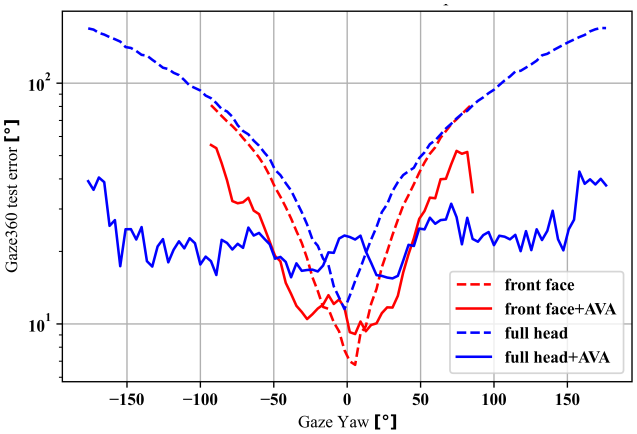

Figure 4: Reduction in gaze error on the Gaze360 test set on jointly training with GazeCapture and AVA-LAEO. The dashed curves are for training with GazeCapture only and the solid ones are for jointly training with GazeCapture and AVA-LAEO. Each curve represents the mean of samples in bins 1.8° wide and the bins with 20 samples or less are discarded. The vertical axis is represented in log scale. Lower is better.

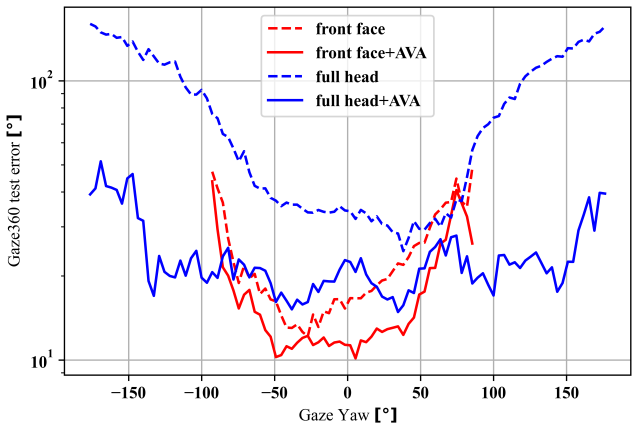

Figure 5: Reduction in gaze error on the Gaze360 test set on jointly training with ETH-XGaze and AVA-LAEO. The dashed curves are for training with ETH-XGaze only and the solid ones are for jointly training with ETH-XGaze and AVA-LAEO. Each curve represents the mean of samples in bins 1.8° wide and the bins with 20 samples or less are discarded. The vertical axis is represented in log scale. Lower is better.

D.1. CMU Panoptic

The CMU Panoptic dataset contains 31 views captured from high-definition cameras within a dome with available accurate body/facial 3D landmark locations and camera intrinsic and extrinsic parameters. This enables us to compute each subject's head position and orientation with respect to any scene camera. Such a convenient setup allows us to quickly gather our own large-scale gaze dataset by lever-

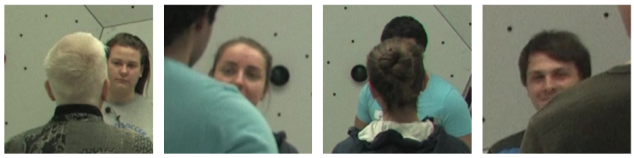

Figure 6: Examples of discarded CMU Panoptic frames from our experiments.

aging the LAEO constraint. However, this dataset does not contain explicit information about the presence or absence of the LAEO activity in video frames. So we use a semi-automatic procedure to label the video frames in it with LAEO activity labels. We use the pre-trained Gaze360 static network [5] to estimate gaze for every subject from multiple frontal views (*i.e.*, if a given face is oriented within $\pm 90^\circ$ of a camera's principal axis). These gaze estimates are then transformed to world co-ordinates and their pairwise cosine distance is computed between every subject pair present in a frame. A pair of gaze vectors for two subjects are assumed to be under LAEO when their angular separation (with one of the vectors being inverted) from each other and the 3D line joining their cyclopean 3D eyes is $< 20^\circ$. A pair of subjects is treated to be in LAEO when at least 4 of its gaze pairs from multiple views are classified as being in LAEO. The nature of the *haggling* activity ensures that only a single pair may ever exhibit LAEO. Frames with none or multiple LAEO pair detections are removed from the analysis.

We experience two corner cases: a) facial features of certain subjects can be blocked from view by another subject in the scene and b) multiple subjects may appear within the same head bounding box (see Fig. 6). To mitigate this issue, we first compute a facial bounding box surrounding a subject's ears, eyes and nose keypoints. Next, we compute a bounding box around every subject's body. Views with facial bounding boxes overlapping with body bounding boxes of other subjects (*i.e.*, with a bounding box IOU score $\geq$0.01) are discarded from the analysis. This in-turn results in missing gaze values in the central $\pm 15^\circ$ gaze pitch and yaw distribution region (see Fig. 3 in the main paper).

D.2. AVA-LAEO

The availability of 3D head poses and landmarks is a vital requirement for computing our LAEO losses. These annotation, however are not available in the AVA-LAEO dataset. We utilize dense 2D-3D correspondence predictions derived from DensePose [4] to fit the SMPL 3D head model to every detected subject within a LAEO annotated frame from the AVA training set [3] with LAEO annotations provided by Marin-Jimenez *et al*. [6]. To improve computational efficiency while deriving these correspondences, we utilize up to 1,000 2D pixels detected by DensePose, which belong to a subject's head. To ensure that every detected facial region is well represented while computing 3D head pose, we uniformly sample 2D pixels based on their distance from the mean 2D head location. However, incorrect head-pose estimates due to incorrect 2D-3D correspondences are inevitable. See Fig. 7 for a positive and a negative example of head pose fitting, where the latter results in noisy gaze labels for the AVA-LAEO dataset.

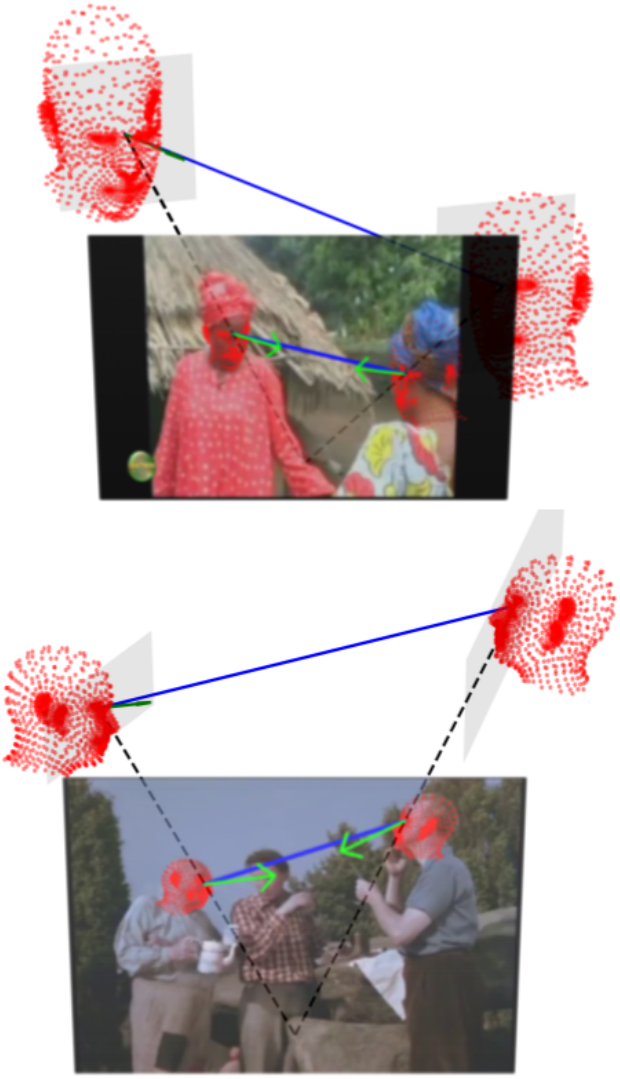

Figure 7: (**Top**) a positive and (**bottom**) a negative example of scene geometry reconstruction from the AVE-LAEO dataset. Notice the incorrect 3D head placement for the right most subject in the bottom example with respect to z depth. The subject on the right is clearly closer to the camera (in terms of z depth) than the subject on the left, but is incorrectly estimated as being further from it. This, in turn, results in noisy 3D gaze labels.

D.3. Reliability of LAEO 3D Gaze Labels

When scene geometry is unknown (*e.g.*, in real-world LAEO datasets), 3D gaze labels derived from LAEO are indeed noisy. We introduce various constraints while training

our system to counter this issue, and show results on both controlled (CMU Panoptic) and in-the-wild (AVA-LAEO) datasets. Yet, as a rough estimate, we compare the angular separation between 3D gaze derived from the approximate scene geometry (described in Sec. 3.3 of the main paper) and its ground truth values using a subset of 3495 images from the CMU Panoptic dataset. On average, we observe a 14.8° gaze label error and an absolute relative depth difference of 0.3 between the ground truth and estimated subject depths when both 2D cyclopean eye points and the subjects' z depths are estimated, and the focal length is assumed to be the largest image dimension. Replacing with accurate focal length reduces gaze label error to 10.1° and using accurate 2D cyclopean eye centers further reduces it to 8.84°. Additionally, the assumption that people look at each others' 3D eye centers introduces $< 4.3^\circ$ gaze error for subjects located > 500mm apart. These label errors are significantly smaller than those encountered in cross-dataset ($\sim 30^\circ$ from [7]) and semi-supervised ($> 25^\circ$ from Fig. 4 of the main paper) training for Gaze360 making LAEO data a reliable source of supervision for 3D gaze learning in physically unconstrained settings.

References

[1] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg. Detecting Attended Visual Targets in Video. In *CVPR*, 2020. 1, 2

[2] K. A. Funes Mora, F. Monay, and J.-M. Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *ACM ETRA*. ACM, Mar. 2014. 1

[3] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *CVPR*, 2018. 5

[4] R. A. Güler, N. Neverova, and I. Kokkinos. DensePose: Dense Human Pose Estimation in the Wild. In *CVPR*, 2018. 5

[5] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. *ICCV*, pages 6911–6920, 2019. 2, 3, 5

[6] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman. Laeo-net: Revisiting people looking at each other in videos. *CVPR*, 2019-June(i):3472–3480, 2019. 5

[7] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*, 2020. 6