Where are *they* looking in the 3D space?
– Supplementary material –

Nora Horanyi[1] Linfang Zheng[1] Eunji Chong[2] Aleš Leonardis[1] Hyung Jin Chang[1]

[1]School of Computer Science, University of Birmingham
[2]Amazon.com, Inc.

{nxh840@alumni, lxz948@student}.bham.ac.uk eunji8703@gmail.com, {a.leonadis,h.j.chang}@bham.ac.uk

References from the paper are **in bold.**

1. The attention target is within the subject's head bounding box

A person cannot look at their face; therefore, the attention target is unlikely to be within the subject's head bounding box. To test our hypothesis, we studied the ground truth head and co-attention bounding box annotations of the VideoCoAttention (VCA) **[8]** joint attention target estimation dataset. We performed an experiment to show how the prediction accuracy changes when we exclude the head bounding box area of the subjects within the scene.

First, we calculated the frequency of the ground truth co-attention bounding box annotation's intersection with the subjects' head bounding boxes (See Table 1a). We found that the intersection occurrence frequency differs among the dataset's training, validation and test set. The highest frequency of intersections was counted in the training set, where 6.42% of the ground truth co-attention annotations intersected with at least one of the head bounding boxes of the frame. While this number is relatively high, we also measured the average AUC within the intersection and found that the AUC score associated with these points was very low. Meaning, that while there exists an intersection between the head and co-attention bounding boxes, the points within this area are not the co-attention target points.

Therefore, we performed an experiment where we excluded the head bounding box area of the image and measured the single and joint attention target estimation accuracy in terms of L2 distance and Out-of-Frame AP (denoted as Inout). In Table 1b, we show that after excluding the points within the head bounding box regions from the attention target estimation, the results improved in the case of both single and joint attention target estimation. Qualitative highlights are shown in Figure 1.

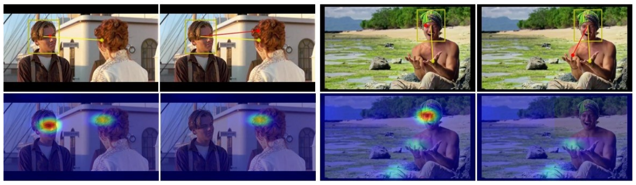

Figure 1. **Qualitative highlights of the head bounding box exclusion experiment.** We show two examples of the recurring error where the model predicts the user to look within their head bounding box at their own face. By excluding the area of their head bounding from the search field the gaze target estimation accuracy improved. The observed subject's head bounding box is highlighted in yellow, the ground truth annotation is marked as a yellow circle or blue dot, and the estimated gaze target estimate is shown as a red circle.

Table 1. **Quantitative comparison on the VideoCoAtt dataset [8] of the SAT and JAT accuracy with and without the head bounding box region.**

(a) Frequency and average AUC score of the head and co-attention bounding box intersections in the train, validation and test sets of VCA.

	Train	Validation	Test
Frequency (%)	6.42	3.67	2.55
Average AUC	0.004	0.002	0.002

(b) Results of the quantitative evaluation on the VCA dataset including and excluding the head bounding box image region for the single and joint attention target prediction.

	Single Attention		Joint Attention	
	L2 dist (px) ↓	Inout (%) ↑	L2 dist (px) ↓	Inout (%) ↑
Include	67.38	54.85	56.48	53
Exclude	**64.58**	**56.48**	**48.70**	**66.53**

2. Failure cases of the existing methods

We conducted a series of experiments on the single attention target estimation to better understand the potential problems that may occur during joint attention target estimation. We used **[3]** as a baseline and took incremental steps to investigate the weaknesses of the state-of-the-art gaze target prediction models. Based on our experiments, we identified the following problems with the existing methods:

- **Bias towards humans:** In Figure 2 a), we show a common mistake of our model. There are many cases in the benchmark datasets where the gaze target of the subject is not another person but an object or the ground in front of them or somewhere between two potential gaze target locations. These cases are especially common in the video dataset, where we observe the user's eye movements frame by frame. During the observed time, the user often shifts their gaze between people or objects. Our analysis showed that our model is biased towards humans, and it is more likely to predict the subject to be looking at another person within the image instead of an insignificant location.

- **Ambiguous annotations:** We studied the annotation of the GazeFollow image dataset, where we have ten annotations for each test image to better understand the ambiguity of the gaze target annotations (example shown in b)). This variation among human annotations originates from the subjective nature of the task. In part b) of the figure, we visualised the ground truth annotations of the selected subject's estimated gaze target location in blue and their average in red in 3D using a prior depth map. We show that annotating the gaze target locations on a 2D image can result in inaccuracies. For example, in b) several ground truth annotations and their average fall behind the subject.

- **Occlusion:** Based on the qualitative results, we identified the two most common causes of failed gaze target estimates, which are:

  - **Occlusion of the subject:** In Subfigure c), we show examples where the occlusion of the subject's face caused the prediction error. Generally, when the model is used to estimate someone's gaze target from the back, the estimate is often different from the ground truth annotations.

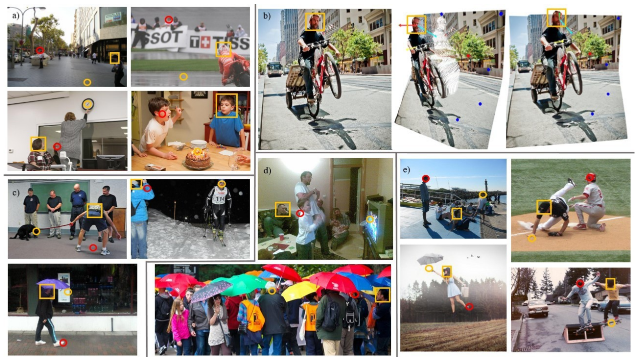

Figure 2. **Visualisation of different failure cases**, including human bias (a), ground truth annotation ambiguity (b), occlusion of the subject (c) and the gaze target (d), and physically impossible estimates (e). The observed subject's head bounding box is highlighted in yellow, the ground truth annotation is marked as a yellow circle or blue dot, and the estimated gaze target estimate is shown as a red circle.

    Note that for these cases, the manually annotated ground truth target locations are not well aligned, meaning that even the human annotators could not agree on the gaze target, which highlights the complexity of this case.

  - **Occlusion of the gaze target:** In other cases, we found that the gaze target selected by the human annotators was occluded by, for example, another person within the image scene (See Subfigure d)). This scenario is not uncommon in the existing benchmark datasets. While it is incorrect to annotate that the subject at the top of Subfigure d) is looking at the television even when the other person is standing in front of him, it makes sense regarding the ongoing activity.

- **Physically impossible estimates:** Finally, we found a common problem where the model predicted the gaze target to be behind the subject or in a physically impossible position. In reality, humans can only look at target points within their field of view, defined as the part of their visual field that can be viewed instantaneously **[S1]**. This error may occur due to an incorrect head pose or gaze direction estimate or when the most probable target is behind the subject.

[S1] Jang, W., Shin, J. H., Kim, M., Kim, K. K. (2016). Human field of regard, field of view, and attention bias. Computer methods and programs in biomedicine, 135, 115-123.

Figure 3. **Single attention target estimation benchmark dataset highlights.** On the left we show images of the GazeFollow image dataset **[24]** and on the right sample image frames of video sequences of the VideoAttentionTarget dataset **[3]**.

3. Dataset and evaluation metrics

3.1. Datasets

3.1.1 Benchmark datasets - Single Attention Target Estimation

GazeFollow dataset. [24] A widely used dataset for predicting the gaze target of the subjects is the GazeFollow benchmark dataset **[24]**, which contains static images. See example images in Figure 3. Amazon Mechanical Turkers annotated the head and gaze locations inside the images of 130,339 people in 122,143 images. 10 different people annotate each image of the test set. The diversity of these annotations well reflects the subjective and complex nature of the gaze target attention estimation task. This dataset does not handle cases when the gaze target is outside the image frame.

VideoAttentionTarget (VAT) dataset. [3] The VideoAttentionTarget video dataset **[3]** is specifically designed for modelling the gaze target in videos. We show an example of randomly sampled image frames of different video sequences in Figure 3 to demonstrate the diversity of the dataset. For each video clip, the annotators provided the head bounding boxes as well as the gaze target of each person with the indication of whether the person was looking outside the video frame.

3.1.2 Benchmark datasets - Social Interaction Detection

Looking At Each Other (LAEO) dataset [19] The video dataset proposed by Marin *et al.* was used to train a model which can analyse one-to-one social interactions between subjects. The primary question they were trying to answer was whether the subjects looked at each other (See examples in Figure 4). The data consist of three types of annotations: a binary label indicating the presence of any pair of people looking at each other, the head bounding boxes of the subjects present at the scene, and, if they exist, the indices of the subjects looking at each other. The dataset is limited to human-human interactions and bounding box-level annotations; no pixel-wise gaze target point is available. The dataset does not extend to cases with more than two participants; therefore, while there are multiple subjects in the scene, joint attention, as we defined it in this work, does not exist in this dataset.

Figure 4. **Social interaction detection benchmark dataset highlights.** On the left we show example image frames of video sequences from the LAEO **[19]** dataset and on the right, we show examples from the VideoCoAttention JAT estimation benchmark dataset **[8]**.

VideoCoAtt dataset (VCA) [8]. A more detailed, larger video dataset proposed by Fan *et al.* was proposed for training models to estimate the joint attention target of the subjects in the video frames. This large-scale, diverse dataset consists of 492,100 from 380 video sequences of 20 different shows. For every frame, they collected the bounding box of joint attention. The attention targets occluded or outside of the frame were not annotated. In addition, they collected the head bounding boxes of the currently engaged subjects within the image frame. The drawbacks of this dataset are that not every person is annotated within the scene, and only one attention target bounding box is identified per image frame. VideoCoAtt dataset highlights are shown in Figure 4.

3.2. Evaluation metrics

In our experiments, we evaluate the performance of the single attention target estimation models on the GazeFollow and VAT benchmark datasets using the following three performance measures: AUC, Distance, and Out-of-Frame AP.

- **AUC:** Each cell in the spatially-discretised image is classified as a gaze target or not. The ground truth comes from thresholding a Gaussian confidence mask centred at the human annotator's target location. The final heatmap provides the prediction confidence score evaluated at different thresholds in the receiver operating characteristic curve (ROC). The area under the curve (AUC) of this ROC curve is reported.
- **Distance:** Pixel-wise normalised L2 distance between the ground truth target location and the pixel of the maximum value in the predicted heatmap.

- **Out-of-Frame AP:** The gaze target estimation model learns a scalar α which quantifies whether the person's focus of attention is located inside or outside the frame, with higher values indicating in-frame attention. The average precision (AP) is computed for the prediction score from the scalar α against the ground truth computed in every frame.

Note that AUC and Distance are computed whenever an in-frame ground truth gaze target (the heatmap always has a maximum). Also, the ten ground truth annotation locations of the GazeFollow dataset were averaged, and the *average L2 distance* was calculated w.r.t. this new ground truth location. We found that the average position was completely off from the actual gaze target in cases where the ground truth annotations disagreed. The *minimum L2 distance* was calculated as the minimum distance from all the ground truth gaze locations. We also show the performance of the annotators (Human performance) across all three measures of the datasets. This is done by comparing annotator predictions in all pairs and averaging them.

The task of Attention Target Detection consists of two subtasks: spatial location prediction and temporal interval detection. To evaluate and compare the performance of the joint attention target prediction models, we used the L2 distance for the localisation task and reported the Prediction Accuracy for the detection task.

- **L2 distance:** Using the predicted joint, joint attention confidence map, we compute the distance between the pixel location of the maximum confidence and the centre of the ground truth bounding box.

- **Prediction Accuracy:** We regarded the given frame with joint attention when the predicted confidence map's maximum value was above a threshold adopted from **[8]**. The Prediction Accuracy is calculated as the percentage of the frames with correct joint attention estimation.

4. Additional qualitative results

We present additional qualitative highlights omitted from the main paper due to space constraints/ The input of the Full and Full-NL proposed models and their variants, the RGB image, the generated prior depth map and the corresponding calculated 3D FOV probability map are shown in the first column. We visualised the generated output heatmap of every variant (Scene only, Scene+depth, Scene+prob), the Full method (Scene+depth+prob), and finally, the gaze target prediction of the Full method. In the target prediction visualisation and the input image, the head bounding box of the observed subjects and the ground truth annotations are marked as yellow, and the estimated gaze target is shown in red.

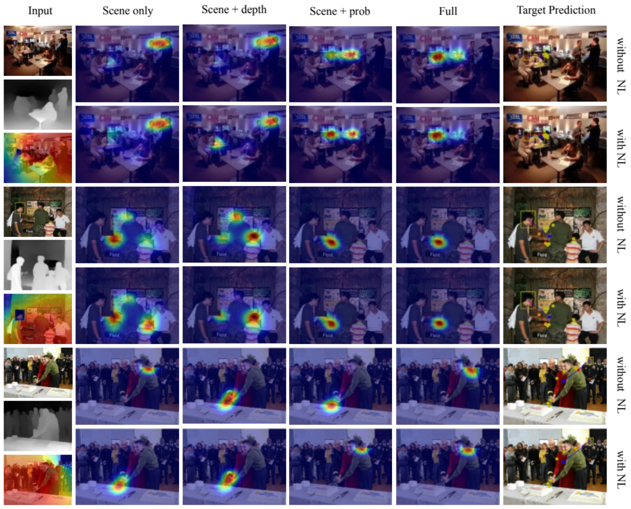

Figure 5. **Qualitative results of ablation study on the GazeFollow SAT benchmark image dataset.**

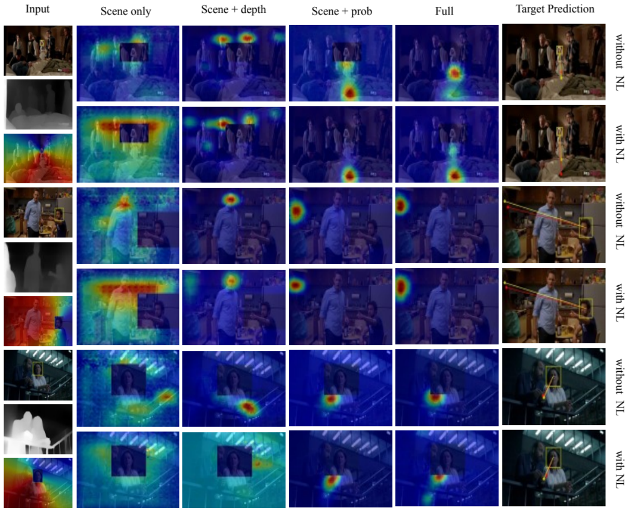

Figure 6. **Qualitative results of ablation study on the GF SAT image, VAT SAT video and the VCA JAT video benchmark datasets, respectively.**

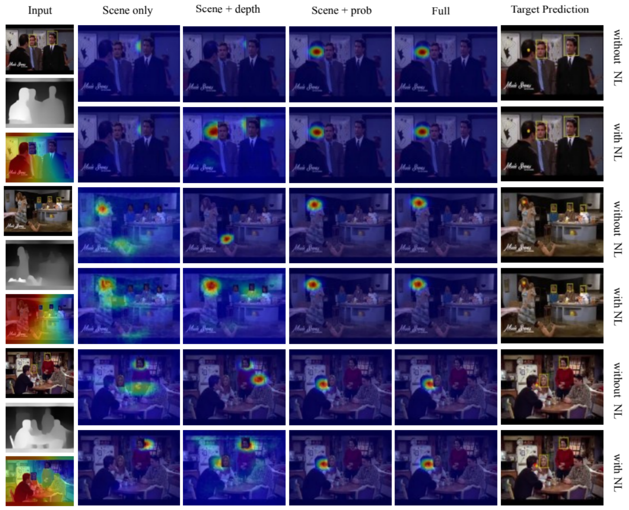

Figure 7. **Qualitative results of ablation study on the VCA JAT benchmark video dataset.**