Published in the 18th IEEE Int. Conference on Automatic Face and Gesture Recognition (FG'24) - Best Student Paper Award, May 2024

A Unified Model for Gaze Following and Social Gaze Prediction

Anshul Gupta*,[1,2] Samy Tafasca*,[1,2] Naravich Chutisilp[2] Jean-Marc Odobez[1,2]
[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne, Switzerland
{agupta, stafasca, odobez}@idiap.ch, naravich.chutisilp@epfl.ch

***Abstract*— Human gaze plays a crucial role in communication and social interaction. Many recent studies have focused on predicting the 2D pixel location of a person's gaze target in an image. However, this approach has limitations when it comes to studying gaze for downstream applications that require analysis of higher-level social gaze behaviors. Previous works have post-processed the predicted 2D gaze target for social gaze prediction, however, we show that this approach is insufficient. Our proposed method jointly predicts the gaze target and social gaze behaviour, explicitly incorporating people interaction for state of the art results on three social gaze tasks - looking at heads, mutual gaze and shared attention. Additionally, we introduce evaluation protocols for these tasks, presenting a promising avenue for future research in gaze behavior analysis.**

I. Introduction

Understanding where and at whom or at what a person is looking is an important task, with applications in consumer behavior analysis [53] [25], human-human [3] or human-robot interaction analysis [23], psychological studies [42] and medical diagnosis [21]. In particular, detecting and understanding social gaze behaviours like mutual gaze [34] or shared attention [13] plays an important role in the development of AI systems with social intelligence. Analysing social gaze is also essential for the assessment of developmental disorders like autism spectrum disorders (ASD) [2]. Given that 1 in 36 children have been identified with ASD [31], having tools to predict social gaze can help decrease time for diagnosis, allowing more children to be tested at lower costs.

Analysing the social gaze of people requires the semantic categorization of their Visual Focus of Attention (VFOA), i.e. are they looking at another person or object. There have been several works in this direction [48][4][43][36], but they often assumed access to frontal views of people and prior knowledge about the 3D structure of the scene (e.g. by relying on multi-camera setups with known geometry), making them unable to generalize to unseen environments. With a focus on the generalization of gaze analysis in arbitrary scenes, Recasens *et al.* [46] introduced the Gaze Following task which aims to predict the 2D pixel location of the gaze target of a person in a scene, using only the RGB image as input. This seminal work has been extended in several ways, by using temporal information [10], additional derived modalities such as depth [23] [17], or allowing the joint processing of multiple people [22][52]. However, the main issue of these works is their focus on predicting the 2D pixel location of people's VFOA, which is problematic since inferring the semantic VFOA is often more relevant to downstream applications. This also means that current performance metrics like euclidean distance between the ground-truth (GT) and predicted gaze target location might not be appropriate to evaluate such systems, as methods with similar performance in pixel space may have very different performance when evaluated semantically.

Fig. 1: Our method accurately predicts social gaze where state of the art Gaze Following methods fail. Row 1 shows sample cases for LAH, and row 2 shows sample cases for LAEO. Orange denotes true positive pairs from our model, while white denotes gaze predictions from [54], which on post-processing leads to incorrect social gaze predictions.

In this paper, moving towards a more semantic analysis, we propose to extend the Gaze Following task to also predict social gaze behaviors in a unified architecture. More specifically, we are interested in three social gaze tasks illustrated in Fig. 2:

- LAH: Looking at Heads, which indicates whether people look at others. This is a fundamental task in social gaze analysis to understand person behaviour like in group conversation analysis where gaze patterns depend on speaking/listening status [18];
- LAEO: Looking at Each Other (i.e. eye-contact, or mutual gaze), which is an important cue in social interaction signaling engagement and attention between individuals [26] and therefore used in the ADOS diagnostic test for autism [29];
- SA: Shared Attention, *i.e.* two or more people looking at

* indicates equal contribution

the same target. This cue contributes to the understanding of the coordination of attention with others and is crucial in the development of children [39] and is also a component measured in ADOS.

Existing work has addressed the social gaze prediction task in two ways. The first approach is the development of models specifically for social gaze prediction. Most methods in this approach aim to predict a specific social gaze task such as shared attention [13] or mutual gaze [34]. Some methods have attempted to predict multiple social gaze tasks [14][8][32], however they suffer from incorrect formulations or slow inference speed. The second approach is post-processing the predicted gaze target from Gaze Following models. For instance, if the predicted gaze targets of two individuals is within a certain distance, they are assumed to share attention [10]. Recently, as Gaze Following models have improved, this approach has been shown to outperform task specific models [16], [10]. However, we show that standard post-processing of Gaze Following models is not enough, probably as it misses scene and interaction understanding relevant for social gaze prediction. On the other hand, our proposed method combines both approaches by 1) leveraging the gaze representations of a recent state of the art Gaze Following model, 2) training the model for social gaze prediction. We further investigate the benefit of explicitly model person-person interactions through a graph module. Our proposed method achieves state of the art results, outperforming the post-processing approach.

Contributions. In this paper we propose three major contributions:

- *Joint Gaze Target and Social Gaze Prediction.* We propose a new model that jointly predicts the gaze target and social gaze behaviour of people in a scene. To this end, we leverage a recently proposed model for Gaze Following, Sharingan [52], that achieves state of the art results on public benchmarks. Sharingan can process the gaze of all people in the scene, unlike typical Gaze Following architectures [10][15][17]. Thus, it naturally extends to social gaze prediction tasks that depend on people interactions. Our model conceptualizes the token-based person representations returned by Sharingan as a fully connected graph, where each person is a node. The nodes are either directly passed to multiple pairwise task heads to predict the social gaze behavior of every edge (i.e. a pair of people), or first updated using a Graph Attention Network to explicitly model people interactions, before being passed to the pairwaise task heads. Extensive experiments shows that our models achieve state-of-the-art or better results compared to standard baseline methods on public benchmarks.
- *Incorporating the Speaking Status.* Although prior VFOA estimation studies have shown improvements using speaking status as auxiliary information [49][43], its exploitation and effectiveness for general scenes has not yet been investigated. Consequently, we extended our model and conducted an initial study to assess whether incorporating video-based inferred speaking status can enhance the looking at heads performance. Results indicate possible performance improvements, but further experiments are required to fully explore the potential benefit of this approach.
- *New Evaluation Protocols*. Tafasca *et al.* [51] recently introduced the LAH task for evaluation on Gaze Following benchmarks. However, their evaluation protocol does not consider the target person. In addition, while people have addressed the SA task [13][10], existing evaluation protocol lack clear definitions. In this paper, we introduce well-defined protocols for both tasks, and will share evaluation scripts.

In the following sections, we describe the related work in this area (Sec. II), our model architecture (Sec. III), experiments (Sec. IV) and conclusion (Sec. VI). All models and evaluation scripts will be made publicly available to facilitate further research in gaze behaviour understanding.

II. Related Work

This paper is about extending a Gaze Following method to also predict social gaze behaviors in a unified architecture. We are specifically interested in Looking at Heads, Looking at Each Other (*i.e.* eye-contact, or mutual gaze), and Shared Attention (*i.e.* two or more people looking at the same target). Next, we review some related works about these tasks.

Gaze Following. Originally introduced in [45], the task of gaze-following aims to predict the pixel-wise target gaze location of a person of interest in the scene. Since attention can be either endogenous (*i.e.* goal-driven, voluntary, top-down) or exogenous (*i.e.* stimuli-driven, reflexive, bottom-up) [7], predicting where a target person is looking requires not only analyzing the person but also understanding the global context of the scene (*i.e.* Where are people located? How are they interacting? Where are the salient objects?). For this reason, typical architectures solve the task by relying on CNN networks with two streams: the first one processes the scene while the second one is more focused on the target person, before feeding into a fusion mechanism to combine information from both branches [45], [10], [15], [17], [22], [23], [28]. Other works brought improvements on several fronts, such as incorporating other relevant modalities like depth [15], [17], [23], [54], [51], [37], pose [17], and objects [20], leveraging scene geometry [19], [23], [15], processing multiple people at once [22], or 1-stage training by jointly predicting the head bounding boxes and gaze targets [55].

Looking at Heads. Tafasca *et al.* [51] recently proposed the Looking at Heads task for evaluating the performance of Gaze Following models. They motivate the importance of the task with applications in child gaze behaviour understanding, for instance identifying child looking at clinician behaviour during an Autism diagnostic test [2]. It also serves as a semantic evaluation of Gaze Following models, providing complementary information to standard Gaze Following metrics, as better performance on these metrics does not always correspond to better LAH performance. We further extend this observation to our other social gaze tasks. Also, their evaluation protocol does not consider the target person, so

looking at any head is considered a positive case. We provide a new evaluation protocol that considers the target looked at person in the metric computation.

Looking at Each Other. Mutual gaze or looking at each other (LAEO) is one of the key components of social communication. The computer vision task of LAEO is defined as the prediction of a binary flag for a pair of people, reflecting whether there is a mutual gaze between them. This task of LAEO in videos was first introduced in [35]. Initially, methods focused on geometric approaches to predict LAEO [35], [44]. Since then, several deep learning based works have attempted to solve this task [11], [6], [34], [33], [9]. More recently, gaze following methods [16], [27] also evaluated performance for LAEO, achieving state of the art results. Note that unlike typical approaches our method has two important differences. Firstly, our method is image based and does not use temporal information. Secondly, we do not predict the head bounding boxes given the existence of highly accurate head detectors [24].

Shared Attention. Shared attention is an important skill, critical for early development, language learning [41], [1], and social cognition [40]. The first work in computer vision to propose the task of predicting shared attention is [13], which also introduced a new public benchmark called the VideoCoAtt dataset. Their method combines the predicted 2D gaze cones of people in the scene with a heatmap of object region proposals. The authors define 2 tasks for their dataset: binary classification to predict whether shared attention occurs in a frame, and location detection to infer the target object of the shared attention. A more recent work [50] improves by directly inferring shared attention from the raw image. Several gaze following methods were also evaluated on this dataset to infer shared attention based on the intensity of the combined predicted gaze heatmaps of the people in the scene [10], [55], [27], and they seem to perform significantly better than methods meant specifically for shared attention detection. This finding suggests that the more general task of gaze-following is useful for other social gaze downstream tasks.

The main problem with the task formulation of [13] is that: 1. it cannot distinguish between multiple shared attention behaviors if they occur in the same frame, and 2. it cannot determine which specific people are sharing attention. Our work takes a different approach that solves both problems: we frame the task as a binary classification between pairs of people. This formulation is more natural and has the added benefit of extending to other social gaze tasks such as mutual gaze.

Multiple Social Gaze Tasks Prediction. To the best of our knowledge, there are only two previous works that attempted to move in the direction of predicting multiple gaze communication patterns simultaneously. The first is Fan *et al.* [14] who proposed a spatio-temporal graph network to hierarchically reason about gaze communication at the atomic-level (ex. shared attention, mutual gaze) and event-level (ex. gaze aversion, joint attention). However, they treat atomic-level gaze behaviours as mutually exclusive which is an incorrect formulation. For example, person A can have mutual gaze with person B, while sharing attention with person C towards B. On the other hand, our framework considers pairs of people in the scene, allowing each pair to be assigned one or more social gaze labels. The second work is that of Chang *et al.* [8], that attempted to address the problem in the previous work by considering a new set of five mutually exclusive static gaze classes for dyadic interactions. However, their method operates on order-dependent pairs of people. Thus, their inference process is very slow as they need to perform $\frac{N!}{(N-2)!}$ forward passes with each image, where N is the number of people in the scene. Instead, our method processes all pairs of people in a single forward pass through a graph module.

Finally, neither of the methods perform gaze following. Our method performs simultaneous gaze following and social gaze prediction, which is particularly important in the case of shared attention to identify the target.

III. Model Architecture

Our architecture is illustrated in Figure 2. It consists of two main components - a Sharingan [52] encoder and the Social Gaze Predictor. First, Sharingan processes the scene and head crops to produce both image and person tokens. These tokens are then used as input to a ViT encoder [12] producing as output a set of person tokens encoding gaze and attention information. The output person tokens are then passed to the gaze target regression decoder and the Social Gaze Predictor. This latter module consists of an optional Graph Attention Network with task specific decoders to predict the different social gaze behaviors. More precisely, the graph processes the person nodes (along with the speaking status when available) to jointly model all person-person gaze (or multimodal) interactions, while the task specific decoders take pairs of updated person node representations. We detail each component below.

A. Sharingan

Sharingan is a hybrid CNN-Transformer that extends the standard ViT [12]. More specifically, the input to the transformer is a set of image tokens and person gaze tokens. The image tokens are obtained following the typical process of converting the image into patches, projecting the patches using a linear layer, and adding positional information. The person gaze tokens, on the other hand, are obtained by feeding the set of input heads and their corresponding head bounding boxes to the Gaze Encoder. Within this module, a CNN backbone processes the head crops to produce gaze embeddings. These embeddings are simultaneously used to predict a 2D gaze vector and produce a gaze token through a linear projection layer. At the same time, a linear projection is used to map the head bounding boxes to the same dimension as the transformer tokens. This location information is then added to the gaze tokens to obtain the final person gaze tokens. All tokens (*i.e.* image tokens and person tokens) are then fed to a ViT encoder composed of several standard

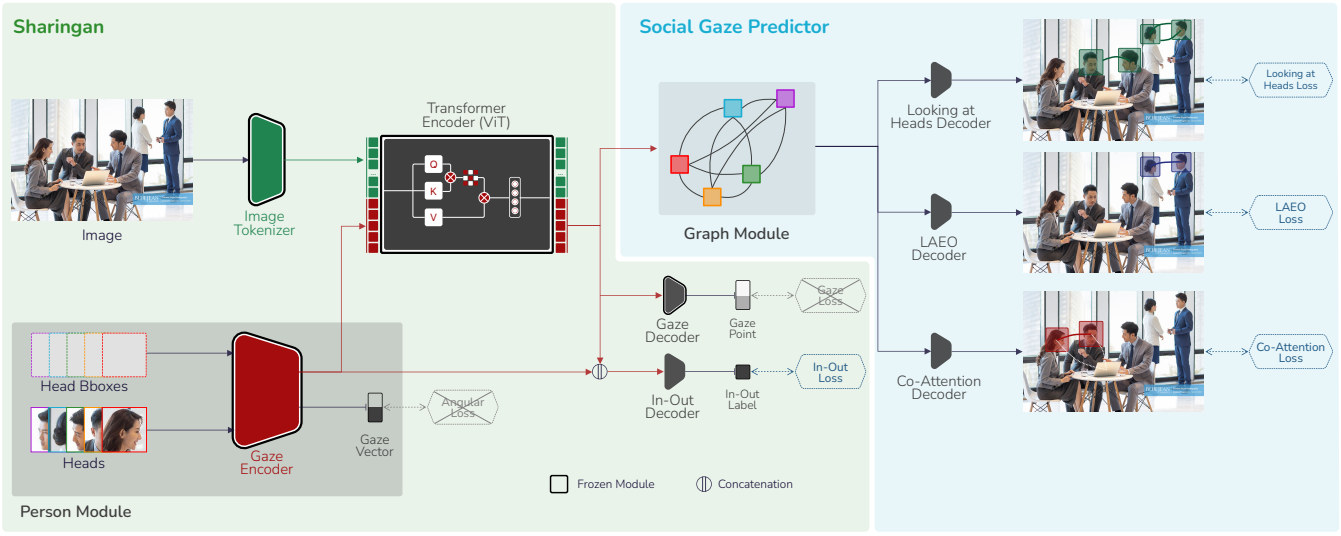

Fig. 2: Illustration of our overall architecture for gaze following and social gaze prediction. It comprises two main components - a Sharingan encoder and the Social Gaze Predictor. The Sharingan encoder first processes the scene and head crops to produce image and gaze tokens. These tokens are then passed to a ViT encoder to generate person tokens encoding gaze and attention information, which are fed to the gaze target regression decoder and the Social Gaze Predictor. The latter module includes a Graph Attention Network with task-specific decoders, which jointly model person-person multimodal interactions and predict different social gaze behaviors. The modules outlined in black are frozen during training.

transformer blocks in order to compute the output tokens. This allows the image tokens and person tokens to interact with each other through self-attention, which updates the person tokens with gaze relevant scene information. The specific output tokens $\mathbf{x}^{\text{out}}$corresponding to the input person gaze tokens are then used by the Social Gaze Predictor and the other decoders to predict pair-wise social gaze, and each individual's 2D gaze point and in vs out of frame gaze label. We refer the reader to the original paper for more details [52].

B. Social Gaze Predictor

We consider two approaches for predicting people's social gaze. In the first approach, we explicitly model people interactions by updating the returned Sharingan person tokens using a Graph Module. The updated tokens are then processed pair-wise by Task Specific Decoders to predict social gaze. In the second approach, we directly pass pairs of Sharingan person tokens to the task specific decoders.

Graph Module. Our goal is to create a model that incorporates all person-to-person gaze interactions, including multimodal interactions when considering speaking status. This facilitates the sharing of visual attention information among groups of people, avoiding unrealistic social gaze configurations. For instance, it can encourage that a person cannot be engaged in a mutual gaze configuration with two other individuals simultaneously by assigning a low weight to such edges.

We achieve the above by relying on a Graph Attention Network [57][5] (GAT) which uses the output person tokens from Sharingan as nodes of a fully connected graph. In general, Graph Neural Networks iteratively update the state of their nodes by aggregating information from neighbouring nodes. In the GAT case, the aggregation mechanism is defined as a weighted average of neighbours in which the weights are computed using an attention mechanism similar to that used in transformers. This feature plays a crucial role in enabling the Graph Module to choose the pertinent subset of individuals for information aggregation.

In a more formal way, the input nodes to the Graph module is the set of N person tokens $\mathbf{x}^{\text{out}}$ from Sharingan. The graph is defined to be fully connected so every node i can attend to every other node j with weight α_{ij}. The output is the set of updated person tokens $\mathbf{x}^{\text{gr}}$.

$$\mathbf{x}_i^{\text{gr}} = \alpha_{ii} W_s + \sum_{j=1, j\neq i}^{N} \alpha_{ij} (W_{\text{gr}} \mathbf{x}_j^{\text{out}}), \tag{1}$$

$$e_{ij} = s^{\text{T}} \text{LeakyReLU}(W_s \mathbf{x}_i^{\text{out}} + W_{\text{gr}} \mathbf{x}_j^{\text{out}}), \tag{2}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{N} \exp(e_{ik})}. \tag{3}$$

Where s, W_s, W_{gr} are learnable parameters. Note that as our graph is fully connected, the Graph Module could also be implemented as a standard Transformer [56]. However, in practice we observe lower performance when doing so, as Transformers have a significantly higher number of parameters and are prone to overfitting on our datasets.

Speaking status. In sufficiently long video segments, a speaking status score be inferred by processing the trajectories of people's head crops [38]. This information is introduced into the model by modifying the person input token. Specifically, the speaking score is transformed into a speaking status embedding $\mathbf{s}$ by linearly projecting the score to the person token dimension. The input to the Graph Module is then defined as the sum of $\mathbf{x}^{\text{out}}$ and $\mathbf{s}$.

Task Specific Decoders. While the graph facilitates the modeling of interactions and is common to all tasks, we rely on task-specific decoders to predict the social gaze behavior for each graph edge. This approach is valid since our behaviors can be defined for every person-to-person pair. In practice, the representation of the edge from persons $\mathbf{p}_i$ and $\mathbf{p}_j$ is given by the concatenated node representations $\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}$ where $\oplus$ denotes the concatenation operation. The task specific decoders are then implemented as Multi-Layer-Perceptrons (MLPs) that process the concatenated node representations to predict the social gaze label. Note that while for LAEO and SA, the prediction is symmetric with respect to to the direction of the edge, for LAH, the prediction depends on the direction of the edge. Accordingly, the decoders are designed as:

$$\mathbf{e}_{i\to j} = E_{LAH}(\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}), \tag{4}$$

$$\mathbf{e}_{i\leftrightarrow j} = E_{LAEO}(\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}), \tag{5}$$

$$\mathbf{c}_{i,j} = E_{SA}(\mathbf{x}_i^{\text{gr}} \oplus \mathbf{x}_j^{\text{gr}}) \tag{6}$$

where $\mathbf{e}_{i\to j}$ denotes whether person i is looking at person j, $\mathbf{e}_{i\leftrightarrow j}$ denotes whether person i and j are looking at each other, and $\mathbf{c}_{i,j}$ indicates whether person i and j are sharing attention.

C. Loss Definition and Model details

Loss. To train the model, we use a combination of the different losses for training:

$$\mathcal{L} = \lambda_{LAH}\mathcal{L}_{LAH} + \lambda_{LAEO}\mathcal{L}_{LAEO} + \lambda_{SA}\mathcal{L}_{SA} + \lambda_{IO}\mathcal{L}_{IO} \tag{7}$$

where $\mathcal{L}_{LAH}$ is the looking at heads loss, $\mathcal{L}_{LAEO}$ is the LAEO loss, $\mathcal{L}_{SA}$ is the shared attention loss and $\mathcal{L}_{IO}$ is the in-out prediction loss, i.e. the loss that defines whether the gaze target of a person is inside the image or not. As each task is a binary classification problem, each loss corresponds to the standard binary cross entropy loss. It is computed for each pair of people and then averaged across all pairs of people in the scene. In principle, we could add as well the target prediction loss. However, as the used datasets for social gaze are small compared to the GazeFollow dataset on which Sharingan is pretrained, we did not observe any improvement in both the target location prediction and social gaze prediction when using it.

Model details. The Sharingan encoder is pre-trained on GazeFollow [46] following the protocol in [52]. For all of our experiments, we keep its weights frozen. The Graph Module is implemented as a GAT with 2 message passing layers and the task specific decoders are implemented as 3 layer MLPs with residual connections. *Inputs.* The scene image and head crops are provided at a resolution of 224×224 to the model. *Speaking Status.* We re-train a state of the art model for speaker detection [38] on the AVA-ActiveSpeaker dataset [47] using just the visual modality. The trained model is then run on our dataset.

IV. Experiments

A. Protocol

Datasets. We leverage the VACATION [14] dataset for joint training, and 3 other datasets for training on individual tasks: VideoCoAtt [13] for shared attention, and UCO-LAEO [34] for looking at each other. As there is no existing dataset for the LAH task, we exploited the annotations in the VideoAttentionTarget dataset [10] to derive the LAH label[1].

VideoAttentionTarget (VAT) [10]. It is a video dataset for Gaze Following, annotated with head bounding boxes, gaze points, and inside vs outside frame gaze for a subset of the people in the scene. It contains 1331 video clips collected from 50 shows on YouTube. The training and test sets contain respectively around 131k and 33k bounding boxes.

To obtain the LAH GT, a highly accurate Yolov5 [24] head detector is used to extract the head bounding boxes of all people in the scene (results were further checked manually for verification). Next, for each annotated person, we check whether their gaze point falls inside any of the detected head bounding boxes. If yes, the label is positive; otherwise it is negative. Overall 69% of the total instances have a positive LAH label.

UCO-LAEO [34]. It was introduced to study the LAEO task (*i.e.* mutual gaze). It is annotated with head bounding boxes, and a label at the head level indicating whether two heads are LAEO. There are $22,398$ images from 4 TV shows, including $6,114$ LAEO pairs from $36,740$ possible head pairs. It is worth noting that we used the Yolo head detector to predict the head bounding boxes of people in "negative" videos (*i.e.* without positive LAEO instances) that are provided without head annotations.

VideoCoAtt [13]. This dataset was introduced in [13] and contains 380 videos of 492k frames. When a shared attention behavior occurs (*i.e.* about 140k frames), the relevant images are annotated with the bounding box of the target object, as well as the head bounding boxes of the people involved. The dataset is split into train/val/test splits of about 250k/128k/114k frames, respectively. Since we also need negative instances, we run a head detector to identify other people in the scene that are not sharing attention. Any pair containing at least one of such people is automatically labeled as a negative instance.

VACATION [14]. This dataset was introduced to study gaze communication in social interactions. It contains annotations for atomic-level and event-level gaze communications. At the atomic-level, it targets both static and temporal gaze behaviours, and at the event-level, it targets temporal gaze behaviours. Given that our model is static, we leverage the static LAEO and SA atomic gaze annotations, and compute LAH annotations following the same protocol as for VAT. It is important to note that the annotation scheme only allows

[1]Doing so for [13] or [34] would produce biased datasets, as annotations are only provided when at least two persons are in a specific configuration (LAEO or SA). Cases with a person looking at another one or in the scene outside these gaze situations are not annotated.

a person to be in a single gaze 'state', so even if they are in LAEO *and* SA, only one of the behaviours is annotated.

Another interesting dataset, GP-static [8], combines clips from VACATION and UCO-LAEO, and annotates them with a new set of social gaze classes. However, this dataset was not available at the time of writing this paper.

Training. During training, we randomly sample up to 6 people in the scene to allow for batch training. At test time we set the batch size to 1 and consider all detected people. The optimization configuration for each task is detailed below.

VideoAttentionTarget. We train the model for 12 epochs with a learning rate of 3e-4 using the AdamW [30] optimizer. The loss coefficients are set as $\lambda_{LAH} = 10$, $\lambda_{IO} = 2$ and $\lambda_{LAEO} = \lambda_{SA} = 0$.

UCO-LAEO & VideoCoAtt. We train the model for 20 epochs with a learning rate of 3e-5 (UCO-LAEO) and 1e-3 (VideoCoAtt) using the AdamW optimizer. For these two tasks, we only use the binary cross-entropy for either the LAEO or co-attention loss and set the remaining losses to 0.

VACATION. We train the model for 20 epochs with a learning rate of 3e-5 using the AdamW optimizer. We set $\lambda_{LAH} = \lambda_{LAEO} = \lambda_{SA} = 1$.

Inference. For the LAH task, a sample is an individual person $\mathbf{p}_i$. At inference, we compute the LAH score for pairs of $\mathbf{p}_i$ with all $\mathbf{p}_j$, and consider the pair $(\mathbf{p}_i, \mathbf{p}_{\hat{j}})$ with the highest predicted score.

$$\mathbf{p}_{\hat{j}} = \text{arg max}_j \ \ \mathbf{e}_{i \to j} \tag{8}$$

For a GT positive case, a prediction is a true positive if $j^* = \hat{j}$ and $\mathbf{e}_{i \to \hat{j}}$ is above the threshold for computing the ROC or Precision-Recall curves. Else the prediction is a false negative. For a GT negative case, a prediction is a true negative if $\mathbf{e}_{i \to \hat{j}}$ is below the threshold for computing the ROC or Precision-Recall curves. Else the prediction is a false positive.

For the LAEO and SA tasks, a sample is a pair of people. A positive case is when there is LAEO and SA respectively, and negative otherwise. Similar to LAH, for LAEO we compute the LAEO score for pairs of $\mathbf{p}_i$ with all $\mathbf{p}_j$, and retrieve the pair with the highest predicted score $(\mathbf{p}_i, \mathbf{p}_{\hat{j}})$. We then set the LAEO score for all other pairs to 0.

$$\mathbf{p}_{\hat{j}} = \text{arg max}_j \ \ \mathbf{e}_{i \leftrightarrow j} \tag{9}$$

$$\mathbf{e}_{i \leftrightarrow j} = 0 \qquad \forall j \neq \hat{j} \tag{10}$$

For the SA task, during inference, we predict the probability of shared attention for each pair of people. For visualization, we perform a post-processing step to get all the disconnected subsets of people sharing attention. In this case, two people will be considered engaged in shared attention if there is a connected path in the graph linking them together.

Metrics. For all tasks, we compute the ROC curve and the associated AUC score, as well as the Precision-Recall curve and the associated AP score.

For the SA task, we also include two metrics to represent the shared attention target localization performance. The first is an L2 distance between the average of the predicted gaze points of all people sharing attention in an image (according to the ground truth) and the center point of the ground truth bounding box of the shared attention target. The second metric, denoted accuracy, measures the number of times the previous average point falls within the target bounding box. We expand the bbox by 5% on each side to account for gaze predictions on the edge. It is worth noting that unlike [13], [10], [55] who compute their metrics per frame, we actually compute our metrics per shared attention instance. Also, our distance is normalized (*i.e.* assuming an image size of 1×1).

B. Tested Models

Post-processing baselines. We evaluate the performance of three Gaze Following models: Chong *et al.* [10], Tonini *et al.* [54], and Sharingan [52] for our social gaze tasks. [10] is a strong baseline model, while [54] is a recent state of the art model. Finally, [52] is the current state of the art for Gaze Following, and the backbone of our architecture.

We post-process the predicted gaze point as follows:

- LAH: For GT positive cases, a prediction is a true positive if the predicted gaze point is inside the GT target head bounding box. Else it is a false negative. For GT negative cases, a prediction is a true negative if the predicted gaze point is not inside any detected head bounding box. Else it is a false positive.
- LAEO: For a pair of people, the prediction is positive if the predicted gaze point for each person is inside the other person's head bounding box. Else it is negative.
- SA: For a pair of people, the prediction is positive if the distance between the predicted gaze points for each person is within a threshold. We define a set of thresholds to obtain ROC and Precision-Recall curves, and the corresponding AUC and AP scores.

Ours-MLP. The output person tokens from Sharingan are fed straight to the task specific decoders without any updates.

Ours-Graph. The output person tokens from Sharingan are first updated using the Graph Module and then passed to the task specific decoders. For a fair comparison, we match the number of parameters for the MLP and graph models.

Speaking. The MLP and graph models augmented with the speaking status information of people in the scene. Denoted by the *spk* subscript.

C. Results

VideoAttentionTarget. Our results for LAH performance on VideoAttentionTarget are summarized in Table I. We see that Ours-Graph provides some improvements over Ours-MLP, with gains of about 1 point for AUC and 0.6 for AP. It also improves over or matches the performance of the Tonini [54] (Prec.=0.900, Recall=0.718) and Sharingan [52](Prec.=0.919, Recall=0.587) post-processing baselines as seen in the precision-recall curve in Fig. 3. However, the Chong [10] baseline (Prec.=0.919, Recall=0.735) slightly outperforms our models.

VideoCoAtt. Our results for SA performance on VideoCoAtt are summarized in Table II. Ours-Graph achieves the best

Model	AUC_{LAH}	AP_{LAH}
Ours-MLP	0.724	0.896
Ours-MLP_{spk}	**0.738**	**0.906**
Ours-Graph	0.733	0.902
Ours-$Graph_{spk}$	0.722	0.893

TABLE I: Results for LAH on VAT. Best results are given in bold.

Model	AUC_{SA}	AP_{SA}
Chong *et al.* [10]	0.695	0.297
Tonini *et al.* [54]	0.715	0.300
Sharingan [52]	0.760	0.301
Ours-MLP	**0.896**	0.594
Ours-Graph	0.890	**0.604**

TABLE II: Results for SA on VideoCoAtt. Best results are given in bold.

Model	AUC_{LAEO}	AP_{LAEO}
LAEO-Net [34]	-	0.795
Doosti [11]	-	0.651
Chang *et al.* [8]	-	0.803
Ours-MLP	**0.986**	**0.957**
Ours-Graph	0.981	0.946

TABLE III: Results for LAEO on UCO-LAEO. Best results are given in bold.

Model	AP_{LAH}	AP_{SA}	AP_{LAEO}
Ours-Graph:LAH	0.924	-	-
Ours-Graph:SA	-	0.213	-
Ours-Graph:LAEO	-	-	0.721
Ours-Graph	**0.935**	**0.222**	**0.741**

TABLE IV: Results on the VACATION dataset. We train for individual tasks (Ours-Graph:{LAH, SA, LAEO }) or all tasks jointly (Ours-Graph). Best results are given in bold.

AP score, and outperforms the post-processing baselines. However, it slightly degrades over Ours-MLP for AUC score. In terms of distance and accuracy, Sharingan [52] (Dist: 0.116, Acc: 0.665) has better performance compared to Chong [10] (Dist: 0.139, Acc:0.579) and Tonini (Dist: 0.127, Acc: 0.452). As the gaze decoder is frozen while training for SA, the scores for Ours-MLP and Ours-Graph remain at distance 0.116 and accuracy 0.665.

UCO-LAEO. Our results for LAEO performance on UCO-LAEO are summarized in Table III. We see that Ours-MLP outperforms other methods to set the new state of the art, with Ours-Graph slightly behind. As most frames in UCO-LAEO contain only 2 people, the graph likely does not bring added benefits. Further, the precision-recall curve in Fig. 4 shows that both models surpass the Chong [10] (Prec.=0.791, Recall=0.832), Tonini [54] (Prec.=0.790, Recall=0.771) and Sharingan [52] (Prec.=0.842, Recall=0.794) post-processing baselines.

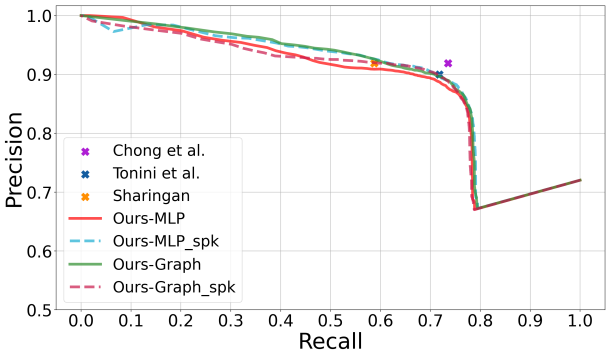

Fig. 3: Prec-Recall curve for LAH on VAT.

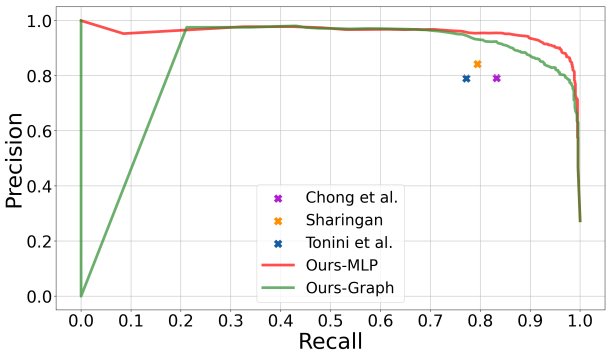

Fig. 4: Prec-Recall curve for LAEO on UCO-LAEO.

VACATION. Our results for LAH, SA and LAEO on VACATION are given in Table IV. We train our model with the Graph Module jointly on all tasks (Ours-Graph), as well as separately on each task (Ours-Graph:{LAH,LAEO,SA }). We observe improvements in performance for all tasks following the joint training, suggesting that it may help in better capturing social interaction semantics.

Speaking Status. Our preliminary results for including speaking status information for LAH are given in Table I. We see an improvement of about 1 point for AUC and AP for Ours-MLP, and a degradation of about 1 point for AUC and AP for Ours-Graph. The improvement in the case of Ours-MLP suggests the potential of leveraging speaking information for improving social gaze inference, however, more research is needed to better capture and incorporate this information.

Gaze Following Metrics vs Social Gaze Metrics. A key observation is that better performance on standard Gaze Following metrics does not always correspond to better social gaze performance. For instance, Tonini [54] reports better Gaze Following performance than Chong [10], but performs worse than it for both LAH and LAEO as seen in Figures 3,4. Similarly, Sharingan [52] reports the best Gaze Following performance but performs worse than Chong [10] for LAH as seen in Figure 3. This can be attributed to the fact that models miss the target person, instead selecting salient items or people nearby. This is not reflected in standard Gaze Following metrics, but is captured through semantic metrics.

Qualitative Results. We provide some qualitative samples from Ours-Graph in Figure 5. We see that it is able to accurately capture social gaze in a variety of situations, including when the face is not visible. It can however fail when it does not capture subtle cues from the eyes that indicate the presence or absence of social gaze.

V. Discussion

Our results in Section IV-C indicate that leveraging a graph module to update the person tokens returned by Sharingan does not necessarily improve performance compared to directly supplying the tokens to the task specific decoders (when matched for parameter count). As Sharingan performs self-attention between all person and scene tokens through the ViT encoder, it may already capture person interactions relevant for social gaze when it is trained for gaze following. Another possible explanation is the small size of the datasets, which may not contain enough diversity for the Graph Module to accurately capture social interactions.

VI. Conclusion

In this paper, we proposed a unified model for Gaze Following and social gaze prediction, focusing on the tasks of looking at heads, mutual gaze and shared attention. Unlike other approaches, our models leverage gaze following representations and are explicitly trained for social gaze prediction. They show improved performance compared to baseline methods on public benchmarks, opening a new direction for more semantic analysis of Gaze Following performance. We also investigated the benefits of joint training on all tasks and observe that it can lead to improvements in performance compared to training on individual tasks. Finally, we showed the potential of incorporating speaking status for improving social gaze prediction through gains in LAH performance in the case of our MLP model. For future work, we plan to further investigate the effectiveness of speaking status information and attempt new ways of multimodal fusion. We also plan to investigate training on larger datasets through new annotations to incorporate greater data diversity.

Acknowledgement. This research has been supported by the AI4Autism project (Digital Phenotyping of Autism Spectrum Disorders in children, grant agreement no. CR- SII5 202235 / 1) of the the Sinergia interdisciplinary program of the SNSF.

References

[1] N. Akhtar and M. A. Gernsbacher. Joint attention and vocabulary development: A critical look. *Language and linguistics compass*, 1(3):195–207, 2007.

[2] A. P. Association. Diagnostic and statistical manual of mental disorders (5th ed.). 2013.

[3] S. Ba and J.-M. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *Transactions on Pattern Analysis and Machine Intelligence*, 33(1), 2011.

[4] C. Bai, S. Kumar, J. Leskovec, M. Metzger, J. Nunamaker, and V. S. Subrahmanian. Predicting the visual focus of attention in multi-person discussion videos. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4504–4510. International Joint Conferences on Artificial Intelligence Organization, 2019.

[5] S. Brody, U. Alon, and E. Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.

[6] G. Cantarini, F. F. Tomenotti, N. Noceti, and F. Odone. Hhp-net: A light heteroscedastic neural network for head pose estimation with uncertainty, 2021.

[7] M. Carrasco. Visual attention: The past 25 years. *Vision research*, 51(13):1484–1525, 2011.

[8] F. Chang, J. Zeng, Q. Liu, and S. Shan. Gaze pattern recognition in dyadic communication. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–7, 2023.

[9] E. Chong, K. Chanda, Z. Ye, A. Southerland, N. Ruiz, R. M. Jones, A. Rozga, and J. M. Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–20, 2017.

[10] E. Chong, Y. Wang, N. Ruiz, and J. M. Rehg. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5406, 2020.

[11] B. Doosti, C.-H. Chen, R. Vemulapalli, X. Jia, Y. Zhu, and B. Green. Boosting image-based mutual gaze detection using pseudo 3d gaze. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1273–1281, May 2021.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] L. Fan, Y. Chen, P. Wei, W. Wang, and S.-C. Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018.

[14] L. Fan, W. Wang, S. Huang, X. Tang, and S.-C. Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5724–5733, 2019.

[15] Y. Fang, J. Tang, W. Shen, W. Shen, X. Gu, L. Song, and G. Zhai. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11390–11399, June 2021.

[16] H. Guo, Z. Hu, and J. Liu. Mgtr: End-to-end mutual gaze detection with transformer. In *Proceedings of the Asian Conference on Computer Vision*, pages 1590–1605, 2022.

[17] A. Gupta, S. Tafasca, and J.-M. Odobez. A modular multimodal architecture for gaze target prediction: Application to privacy-sensitive settings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5041–5050, 2022.

[18] S. Ho, T. Foulsham, and A. Kingstone. Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PloS one*, 10(8):e0136905, 2015.

[19] Z. Hu, D. Yang, S. Cheng, L. Zhou, S. Wu, and J. Liu. We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*, 2022.

[20] Z. Hu, K. Zhao, B. Zhou, H. Guo, S. Wu, Y. Yang, and J. Liu. Gaze target estimation inspired by interactive attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8524–8536, 2022.

[21] L. Isaac, J. Vrijsen, M. Rinck, A. Speckens, and E. Becker. Shorter gaze duration for happy faces in current but not remitted depression: Evidence from eye movements. *Psychiatry Research*, 218(1-2):79–86, 2014.

[22] T. Jin, Z. Lin, S. Zhu, W. Wang, and S. Hu. Multi-person gaze-following with numerical coordinate regression. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[23] T. Jin, Q. Yu, S. Zhu, Z. Lin, J. Ren, Y. Zhou, and W. Song. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence*, 113:104924, 2022.

[24] G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, NanoCode012, Y. Kwon, TaoXie, K. Michael, J. Fang, imyhxy, Lorna, C. Wong, Z. Yifu, A. V, D. Montes, Z. Wang, C. Fati, J. Nadar, Laughing, UnglvKitDe, tkianai, yxNONG, P. Skalski, A. Hogan, M. Strobel, M. Jain, L. Mammana, and xylieong. ultralytics/yolov5: v6.2 - YOLOv5 Classification Models, Apple M1, Reproducibility, ClearML and Deci.ai integrations, Aug. 2022.

[25] N. Kim and H. Lee. Assessing consumer attention and arousal using eye-tracking technology in virtual retail environment. *Frontiers in Psychology*, 12:665658, 2021.

[26] K. Kompatsiari, F. Ciardo, V. Tikhanoff, G. Metta, and A. Wykowska. It's in the eyes: The engaging role of eye contact in hri. *International Journal of Social Robotics*, 13:525–535, 2021.

[27] P. Li, H. Lu, R. W. Poppe, and A. A. Salah. Automated detection of joint attention and mutual gaze in free play parent-child interactions.

Fig. 5: Sample of qualitative results from the test sets. The left block corresponds to SA (*i.e. VideoCoAtt* dataset). The middle block corresponds to LAEO (*i.e. UCO-LAEO* dataset). The right block corresponds to LAH (*i.e. VideoAttentionTarget* dataset). Shapes of the same color denote a relationship (*e.g.* shared attention). The colors orange and pink denote different instances of true positives, while the dashed white corresponds to false negatives and dashed red to false positives. For LAH, the source person is represented by a bounding box, while the target head is denoted by a circle.

In *Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 374–382, 2023.

[28] D. Lian, Z. Yu, and S. Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018.

[29] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30:205–223, 2000.

[30] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[31] M. J. Maenner, Z. Warren, A. R. Williams, E. Amoakohene, A. V. Bakian, D. A. Bilder, M. S. Durkin, R. T. Fitzgerald, S. M. Furnier, M. M. Hughes, et al. Prevalence and characteristics of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2020. *MMWR Surveillance Summaries*, 72(2):1, 2023.

[32] M. Malik and L. Isik. Relational visual representations underlie human social interaction recognition. *Nature Communications*, 14, 2023.

[33] M. J. Marín-Jiménez, V. Kalogeiton, P. Medina-Suárez, , and A. Zisserman. LAEO-Net++: revisiting people Looking At Each Other in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[34] M. J. Marin-Jimenez, V. Kalogeiton, P. Medina-Suarez, and A. Zisserman. Laeo-net: Revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[35] M. J. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106:282–296, 2014.

[36] B. Massé, S. Ba, and R. Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2711–2724, 2017.

[37] Q. Miao, M. Hoai, and D. Samaras. Patch-level gaze distribution prediction for gaze following. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 880–889, 2023.

[38] K. Min, S. Roy, S. Tripathi, T. Guha, and S. Majumdar. Learning long-term spatial-temporal graphs for active speaker detection. In *European Conference on Computer Vision*, pages 371–387. Springer, 2022.

[39] C. Moore, P. J. Dunham, and P. Dunham. *Joint attention: Its origins and role in development*. Psychology Press, 2014.

[40] P. Mundy and L. Newell. Attention, joint attention, and social cognition. *Current directions in psychological science*, 16(5):269–274, 2007.

[41] P. Mundy, M. Sigman, and C. Kasari. A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and developmental Disorders*, 20(1):115–128, 1990.

[42] S. Muralidhar, R. Siegfried, J.-M. Odobez, and D. Gatica-Perez. Facing employers and customers: What do gaze and expressions tell about soft skills? In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*, pages 121–126. ACM, 2018.

[43] K. Otsuka, K. Kasuga, and M. Kohler. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 191–199, 2018.

[44] C. Palmero, E. A. van Dam, S. Escalera, M. Kelia, G. F. Lichtert, L. P. Noldus, A. J. Spink, and A. van Wieringen. Automatic mutual gaze detection in face-to-face dyadic interaction videos. In *Proceedings of Measuring Behavior*, volume 1, page 2, 2018.

[45] A. Recasens*, A. Khosla*, C. Vondrick, and A. Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution.

[46] A. Recasens, C. Vondrick, A. Khosla, and A. Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017.

[47] J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, S. Ramaswamy, A. Stopczynski, C. Schmid, Z. Xi, et al. Ava active speaker: An audio-visual dataset for active speaker detection. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4492–4496. IEEE, 2020.

[48] S. Sheikhi and J.-M. Odobez. Combining dynamic head pose–gaze mapping with the robot conversational state for attention recognition in human–robot interactions. *Pattern Recognition Letters*, 66:81–90, 2015.

[49] R. Siegfried and J.-M. Odobez. Visual focus of attention estimation in 3d scene with an arbitrary number of targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3153–3161, 2021.

[50] O. Sumer, P. Gerjets, U. Trautwein, and E. Kasneci. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3327–3336, 2020.

[51] S. Tafasca, A. Gupta, and J.-M. Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[52] S. Tafasca, A. Gupta, and J.-M. Odobez. Sharingan: A transformer-based architecture for gaze following. In *arXiv*, 2023.

[53] H. Tomas, M. Reyes, R. Dionido, M. Ty, J. Mirando, J. Casimiro, R. Atienza, and R. Guinto. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3125–3133, 2021.

[54] F. Tonini, C. Beyan, and E. Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 420–431, New York, NY, USA, 2022. Association for Computing Machinery.

[55] D. Tu, X. Min, H. Duan, G. Guo, G. Zhai, and W. Shen. End-to-end human-gaze-target detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2192–2200. IEEE, 2022.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[57] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.