A Modular Multimodal Architecture for Gaze Target Prediction: Application to Privacy-Sensitive Settings

Anshul Gupta, Samy Tafasca, Jean-Marc Odobez
Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne, Switzerland
{agupta, stafasca, odobez}@idiap.ch

Abstract

Predicting where a person is looking is a complex task, requiring to understand not only the person's gaze and scene content, but also the 3D scene structure and the person's situation (are they manipulating? interacting or observing others? attentive?) to detect obstructions in the line of sight or apply attention priors that humans typically have when observing others. In this paper, we hypothesize that identifying and leveraging such priors can be better achieved through the exploitation of explicitly derived multimodal cues such as depth and pose. We thus propose a modular multimodal architecture allowing to combine these cues using an attention mechanism. The architecture can naturally be exploited in privacy-sensitive situations such as surveillance and health, where personally identifiable information cannot be released. We perform extensive experiments on the GazeFollow and VideoAttentionTarget public datasets, obtaining state-of-the-art performance and demonstrating very competitive results in the privacy setting case. [1]

1. Introduction

As an indicator of attention, gaze is an important cue which can reveal considerable information about a person's behavior or state of mind. In this regard, identifying the gaze of people in visual data finds applications in many domains, like in the retail industry to understand consumer behaviour [37], in sociology for assessing social gaze behaviours such as shared attention [8], or in human-robot interaction for communication analysis [32].

In recent years, there has been an increased body of work devoted to gaze analytics. One research direction focuses on improving the raw gaze prediction, defined as the 3D angular values representing the 3D line of sight. These methods typically use the eyes [40] or the face of a person [17] as input. Another research line addresses the identification of the visual focus of attention (VFOA), *i.e.* the visual target a person is looking at [3, 26, 32]. Such a task is challenging, as it requires not only to capture the body, head and potentially the eyes of the person of interest to infer their attention, but also the understanding and monitoring of the scene containing the gaze targets. Due to this constraint, traditional methods usually perform the task in fixed environments relying on multi-camera set-ups [3, 26] or prior knowledge of scene location [32]. This creates challenges for applying the methods and models to unseen environments.

Figure 1. Sample images where depth and pose information can be useful to infer the gaze target. Left: the pose can indicate that people are interacting, while depth helps to rule out salient objects in the background. Middle: manipulation activities where knowledge of the hands can be useful. Right: depth allows to filter out the potential face candidates in the back.

With a focus on generalization, Recasens et al. [30] proposed to address the VFOA task using a single image, by formulating it as the prediction of the image 2D location of the gaze target looked at by a person of interest in the image. They proposed a model that implicitly learns what the salient elements in a scene are, and how to combine them with the attention evidence obtained from the gaze information inferred from the person's head crop image. Relying on a large annotated dataset, good results were obtained, with the advantage that such an approach can, in theory, be applied to any arbitrary scene. This work has also been extended in many ways [6, 12, 16, 18, 23, 43]. In this paper, we address this 2D target prediction task and investigate differ-

[1] We plan to release the code after clean-up.

ent factors which can contribute to the success and understanding of the inference process, as motivated below.

Motivations. Several cues modulate gaze following behavior in humans, such as saliency and social context [33]. Transposing this idea to images, we argue that predicting the gaze target of a person in an image can benefit from leveraging such information. This includes inferring the general gaze direction of the person, and identifying the salient items (potential VFOA targets) located in their field of view, like objects and faces. To further remove ambiguities, humans usually rely on priors about gaze behaviours, which depend on understanding the 3D structure of the scene (to check visibility factors) as well as the current context (task performed, ongoing interactions, intentions, past actions, . . .). While the information to reach this level of understanding and apply the right prior for inferring the gaze is directly available in the image, one may wonder whether explicitly providing visual cues and modalities would ease and improve the inference of gaze target.

In this regard, we study the use of explicit depth and pose information for improving attention inference, as illustrated in Fig. 1. Indeed, depth information gives the model an idea about object shapes and the 3D scene structure. It allows for understanding whether a person is looking to the foreground or background, and can help resolve ambiguities along the line of sight when the depth does not match. On the other hand, pose provides accurate information about the locations of body parts related to attention such as hands and faces which are common gaze targets during interactions and manipulation activities. Pose also provides information about the physical state and activity of the person(s), which can help decide which categories of potential gaze targets the network needs to focus on.

Using images with visible faces can be an issue in privacy-sensitive scenarios, either when obtaining training data, or at inference time. Surveillance and health are typical applications. For instance, it has been shown that reduced eye contact and shared attention are early warning signs for autism in children [44]. However, due to their sensitive nature, raw videos (even with faces blurred since we are interested in gaze) are not available for public access, making it difficult to develop and test models. On the other hand, pose and depth data do not contain identifying information, so they can be shared. Hence, we develop models which use only pose and depth data, and evaluate their performance in this paper.

We also study the benefit of other technical elements. The first one is resolution. Indeed, gaze target localization is a dense prediction task similar to pose landmarks estimation: our output is a heatmap corresponding to the probability of a point being the gaze target. Current approaches [7, 9, 18] typically use ResNet style architectures where the spatial resolution of features is greatly reduced before being upsampled again for the final prediction. Instead, we adopt a Feature Pyramid Network approach [19] which includes skip connections during the upsampling process to preserve spatial information and demonstrate improved results. Secondly, in current methods, gaze information is often merged with the input image content to infer the gaze of a person. This early fusion requires the full image to be (re)processed for each person. We investigate whether a late fusion approach can be adopted (fusing gaze information with feature maps) and show that it does not achieve the same level of performance.

Approach and contributions. In summary, we address the gaze target location prediction task and make the following contributions:

- we propose a modular multimodal gaze prediction architecture with end-to-end training and an attention scheme to combine the saliency features of image, pose and depth modalities;
- we investigate the use of only pose and depth information in our setting to allow its usage in privacy-sensitive settings;
- we propose the use of a Feature Pyramid Network regression scheme and show that preserving spatial information is important due to the nature of the task.

We conduct experiments on the GazeFollow [30] and VideoAttentionTarget [7] public benchmarks, and show that we obtain state-of-the-art results using all modalities, and that competitive results can be obtained using only depth and pose information which is of interest for privacy-sensitive applications.

2. Related Work

This paper relates mainly to the problems of gaze target prediction and, to a minor extent, data anonymization. Below is a review of works in these topics.

2.1. Gaze Target Prediction

When accurate gaze trackers were not available, VFOA was often inferred from head pose using behavioral models [35]. Inference mechanisms like GMM, HMM, or Dynamical Bayesian Networks [1, 27] were used to estimate the VFOA directly from the head pose, and potentially using other contextual information [10] which can act as priors on the VFOA like the speaking status, the speech semantic content [26, 32], or modeling interactions and the joint VFOA of all participants [2, 22]. Nevertheless, with the recent improvement of gaze estimation, even simple frame-based geometrical models were shown to be effective to estimate VFOA [42]. Recent models used deep networks such as CNNs and RNNs, resulting in further improved performance [34]. However all these methods typically rely on some prior knowledge about the scene structure and hence can not generalize to arbitrary settings.

To address generalization, Recasens et al. [29] formulated the problem as the inference of the 2D image position corresponding to the location of the scene target a person (in the image) is looking at. They proposed a CNN model combining the information from two branches, a saliency branch which processes the scene and a gaze branch analysing the head crop of the person of interest. Most models that followed relied on a similar two branch architecture. For instance, Chong et al. [6] extended the model to also predict whether a person is looking inside or outside the frame. Lian et al. [18] predicted a 2D gaze vector from the gaze branch and used it to generate explicit gaze cones which were then concatenated along with the input image for inference. We follow a similar idea to generate a gaze cone which is concatenated with each modality, and we investigate the privacy-sensitive setting. Drawing inspiration from works in human pose estimation, Zhao *et al*. [43] proposed an interesting method which learned to predict the line of sight as well as infer the attention 'landmark', and demonstrated improved results over the other baselines. Other works proposed to process multiple people together [16] or use temporal information [7] using an LSTM module at the bottleneck, but in the latter case, results were not improved much compared to the frame-based case.

There has also been some work exploring the use of multimodal information. Guan et al. [12] used the pose of the person of interest to supplement the gaze branch in cases where the face is not visible. In the approach of Nan et al. [23], authors aim to merge (task driven) top-down attention with bottom-up features (flow and pose) to derive the gaze target. While they perform a similar late fusion of features across modalities, our fusion mechanism operates at a much higher resolution and relies on an attention mechanism. In addition, their overall method (with top-down features) is quite different and was applied in a specific setting. Fang et al. [9] used depth to potentially disambiguate attention targets by inferring whether a person is looking toward their foreground or background, obtaining the best results reported so far on the GazeFollow and VideoAttentionTarget datasets. Recently, Hu et al. [14] used depth information to perform gaze target prediction in 3D. As far as we are aware, ours is the first work to use both pose and depth information, and to study the privacy preserving situation.

Finally, there are other works addressing tasks related to gaze following. This includes predicting gaze target objects [37], detecting whether two people are looking at each other [21] or recognizing shared attention behavior [8]. However, due to their aims, these works differ substantially from the study we conduct here.

2.2. Data Anonymization

The typical approaches for anonymizing face data include techniques such as blurring and pixelation [5] [11]. However, these methods may not remove privacy-sensitive information [24, 25] and may instead remove critical information for the downstream task. More recent methods use generative adversarial methods [15] to alter the face. However, these methods are not suited for our task as changes to facial features can affect the gaze information. Instead, we propose the use of pose information (facial landmarks) to predict the gaze direction. This removes identifying information while still giving us a good approximation of a person's gaze direction. This was recently demonstrated by Belkada *et al*. [4], who showed that the head and body pose alone could be used to predict the "eye contact" of people with a camera sensor placed on a car. Our results further confirm this hypothesis in more general scenes.

3. Model Architecture

3.1. Approach overview

An overview of our system is illustrated in Figure 2. It takes as input an image or a video frame, a set of derived modality images, and the head bounding box of a target person. The output is a gaze heatmap $\mathbf{H}$ where the location of the maximum value corresponds to the desired gaze prediction $\mathbf{p}_{gaze}$.

Our network architecture consists of 3 modules. The first one is a *Human-Centric module* whose goal is, given the head crop of a person, to predict a gaze cone representing their visual field of view, *i.e*. the set of pixel locations where the person might be looking. The second one is a *Scene-Centric module* which is fed the image, the person's location (head mask) and the gaze cone in order to generate a feature saliency map $\mathbf{F}$ highlighting possible gaze target locations. The last one is a *Prediction module* comprising two heads: one for inferring the gaze heatmap, and the second one for predicting whether the gaze target is located within the frame. These components are detailed below.

3.2. Human-Centric Module

In this module, a sub-network $\mathcal{G}$ takes the head crop image $\mathbf{I}_{head}$ of the target person as input and predicts a normalized 2D gaze vector $\mathbf{g}_{2D} = \mathcal{G}(\mathbf{I}_{head})$. This gaze vector is used by a *gaze cone generator* to produce a gaze cone image $\mathbf{I}_{co}$. Finally, the gaze cone image is concatenated with the binary head mask of the target person $\mathbf{I}_{mask}$ and passed to the Scene-Centric branch for further processing.

Gaze Cone Generator. The gaze cone is a way to modulate the image information appearing in the gaze direction of the person. It is encoded as an image in order to be consistent with the rest of the architecture. The gaze cone generator produces $\mathbf{I}_{co}$ by drawing a cone from the subject's eyes location $\mathbf{p}_{eye}$ (i.e. eye mid-point if available from the pose modality; otherwise, using a prototypal location in the head bounding box) along the direction of $\mathbf{g}_{2D}$. To account for

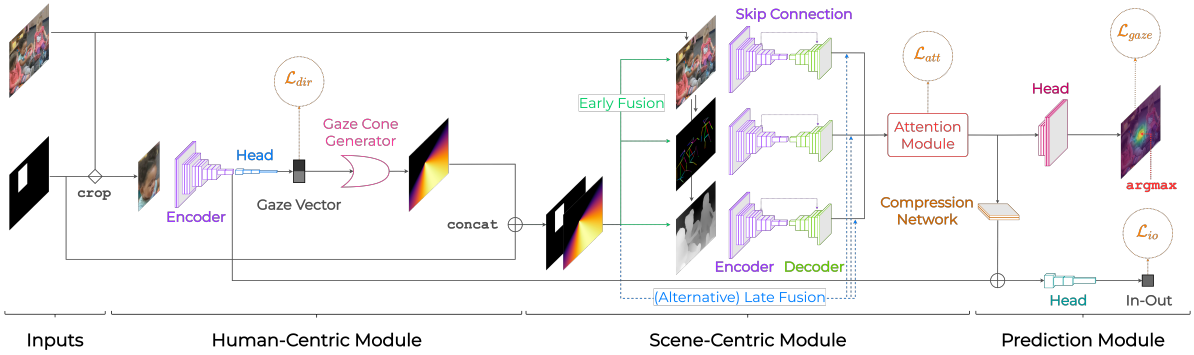

Figure 2. Overview of our proposed architecture. Given an input image and a target subject's head location, we first extract depth and pose maps from the image using off-the-shelf pre-trained models. Next, the Human-Centric module takes the person's head crop as input and predicts a 2D gaze vector which is used to generate a gaze cone image. Then, the Scene-Centric module processes the original image, the depth image and the pose map in order to produce modality saliency feature maps (using modality-specific encoder-decoder feature extractors) which are fused by an Attention module. The resulting saliency map is used by the Prediction module to regress a gaze heatmap, and optionally predict an in-vs-out gaze classification score.

uncertainties in gaze prediction, the cone has an aperture of α_{co} (set to π in practice), and the intensity decays the farther we are from the gaze direction angle-wise. Specifically, the value at each pixel location $\mathbf{p}$ is scored according to the cosine similarity between the predicted gaze vector and the eye-to-target direction. (see example in Fig. 2). Formally:

$$\begin{aligned} \forall \mathbf{p} = (i,j) \text{ where } (i,j) \in [1,\ w] \times [1,\ h], \\ \mathbf{I}_{co}(\mathbf{p}) = \max\left(0, \cos(\mathbf{g}_{2D}, \mathbf{p} - \mathbf{p}_{eye})\right) \end{aligned} \tag{1}$$

Note that given this definition, the gaze cone generator is differentiable, allowing to train our architecture end-to-end.

3.3. Scene-Centric Module

In the Scene-Centric module, the input image $\mathbf{I}$ is first transformed using different networks (see implementation details) into a set of modality images $\mathbf{I_m}$, where $m \in \{raw, pose, depth\}$ and $\mathbf{I_{raw}} = \mathbf{I}$ by definition. These modalities are passed through feature extractors to produce feature maps, which are then fused using an attention mechanism to create a single combined feature map.

Feature Extractors. A set of modality-specific feature extractors $\mathcal{F}_m$ are used to compute feature maps $\mathbf{F}_m$ to encode the person-specific salient regions of the scene according to the input modality. Thus, each feature extractor $\mathcal{F}_m$ processes its corresponding modality $\mathbf{I_m}$ concatenated with the output of the Human-Centric module, so that we have:

$$\mathbf{F}_m = \mathcal{F}_m(\mathbf{I_m}, \mathbf{I}_{co}, \mathbf{I}_{mask}) \tag{2}$$

Note that the concatenation can be seen as an early fusion scheme, whereas an alternative (so far less successful) approach consists in fusing the Human-Centric module information later at the feature level (see late fusion experiments). While multiplication is a more straightforward way to fuse the modality image and the gaze cone, in practice it produced worse results, probably because it performs a hard decision based on potentially inaccurate gaze direction predictions. This is particularly the case when the subject's head is facing backwards and the gaze vector is more difficult to estimate. Concatenation on the other hand, allows the model to make that decision later in the processing.

Regarding the network, we used a typical image-to-image approach, relying on an encoder-decoder architecture. However, in contrast to previous works which simply upsample the lowest resolution representation produced by the encoder [7, 18], we used skip connections from different intermediate representations (at different resolutions) to their corresponding decoder representations in the style of a Feature Pyramid Network [19]. This architectural choice aims to retain information from higher resolution representations, which is important in dense prediction tasks, and further evidenced by our experiments.

Attention Module. Its goal is to perform a soft-selection of the most appropriate input modality given the scene. It takes as input the set of feature maps $\mathbf{F}_m \in \mathbb{R}^{w \times h \times d_m}$ and produces a single combined feature map $\mathbf{F}$, which we use to predict the outputs. Concretely, it performs fours steps:

1. Each feature map $\mathbf{F}_m$ is passed through a modality-specific convolution layer to produce a transformed feature map $\mathbf{T}_m$.
2. Each map $\mathbf{T}_m$ is passed through a network $\mathcal{A}_m$ consisting of three strided convolution layers followed by a global max pooling to generate an embedding vector e_m. All embeddings are then concatenated to form the global embedding e.
3. The global embedding is passed through a projection layer P followed by a softmax operation to get the attention weights: $\{w_m\} = \text{softmax}(P(e))$.
4. Finally, the output is computed as the weighted sum of the transformed feature maps: $\mathbf{F} = \sum_m w_m \mathbf{T}_m$.

This loosely resembles the self-attention mechanism in a transformer [38]: the transformed feature maps $\mathbf{T}_m$ act as the values, whereas the attention weights w_m simulate a dot product between an implicit query and a set of keys.

In addition, this attention mechanism allows us to use a variable number of modalities during inference because the model can simply assign a weight of 0 when a modality is absent. To encourage this behaviour, we perform modality dropout during training, *i.e*. we randomly provide a white noise image instead of the dropped modality, and use an attention loss for supervision (see Section 3.5).

3.4. Prediction Module

This module uses the feature map to predict the quantity of interest: a gaze heatmap $\mathbf{H}$, and a binary In-Out flag o indicating whether the gaze target is inside or outside the image. It comprises two parts, which are explained below.

Gaze prediction head. The gaze target heatmap $\mathbf{H}$ is regressed from the combined feature map $\mathbf{F}$ using a prediction decoder $\mathcal{R}$ that consists of an analytic upsampling followed by a set of convolution layers:

$$\mathbf{H} = \mathcal{R}(\mathbf{F}) \tag{3}$$

The location where the heatmap is maximal is then used as the gaze target prediction.

In-Out prediction head. In general, we want to predict whether the person is looking at a scene location which is visible in the image or not. This is important as we do not want to use the gaze target prediction when a person is looking outside the frame. To accomplish this, we attach an In-Out network prediction head $\mathcal{O}$ which takes as input the feature map $\mathbf{F}$ resulting from the attention step as well as a gaze embedding e_{gaze} coming from the human centric module (see Fig. 2):

$$o = \mathcal{O}(\mathbf{F}, e_{gaze}) \tag{4}$$

More precisely: first, the map $\mathbf{F}$ is passed through a network having the same architecture as $\mathcal{A}_m$ to produce a scene embedding e_{scene} which is concatenated with the gaze embedding e_{gaze} and fed into an In-Out predictor consisting of 2 linear layers followed by a sigmoid activation.

3.5. Loss

The complete model is trained end-to-end using a combination of four losses:

1. **Gaze loss** $\mathcal{L}_{gaze}$. It measures the error in gaze location prediction, which is done by computing the pixel-wise L2 loss between the predicted heatmap $\mathbf{H}^{pred}$ and the ground truth gaze target heatmap $\mathbf{H}^{gt}$, defined as a gaussian blob centered on the ground-truth location.
2. **Gaze direction loss** $\mathcal{L}_{dir}$. The goal of this loss is to better constrain the learning of the Human-Centric module. This is achieved by maximizing the cosine of the angle between the predicted 2D gaze vector $\mathbf{g}_{2D}$ and the ground truth vector $\mathbf{g}_{2D}^{gt}$, which we derive from the ground-truth gaze point.
3. **In-Out loss** $\mathcal{L}_{io}$. We use a standard binary cross entropy loss to measure whether a person is looking inside or outside the image frame.
4. **Attention loss (modality drop)** $\mathcal{L}_{att}$. This loss aims to supervise the Attention module (Section 3.3). The idea is to push the attention weight w_m of a dropped modality m towards 0. Formally, the loss is defined as: $\mathcal{L}_{att} = \sum_m w_m . \mathbb{1}_{m \in \text{dropped}}$, where $\mathbb{1}$ is an indicator variable and 'dropped' is the list of dropped modalities.

The final loss is a linear combination of the four losses:

$$\mathcal{L} = \lambda_{gaze}\mathcal{L}_{gaze} + \lambda_{dir}\mathcal{L}_{dir} + \lambda_{io}\mathcal{L}_{io} + \lambda_{att}\mathcal{L}_{att} \tag{5}$$

3.6. Implementation Details

Modality extraction. The pose maps are extracted using HRFormer [41], a mix between HRNet [39] and the standard transformer architecture. On the other hand, the depth maps are extracted using MiDaS [28], a strong monocular depth estimator. Pose maps are represented as RGB images of skeletons of the people in the frame, where different colors denote the different limbs and keypoints.

Feature extraction networks $\mathcal{F}_m$**.** The feature extractors in the scene branch use backbones chosen from the EfficientNet family [36] because of their ability to scale capacity and expressive power without incurring a significant cost in terms of size. Specifically, the image encoder is an EfficientNet-B1 (7.8M parameters) while the depth and pose encoders use an EfficientNet-B0 (5.3M parameters). Input modalities are resized to 224×224 and fed to the EfficientNet backbones which compute different intermediate feature representations at resolutions between 56×56 and 7×7. These are used in the residual connections of the Feature Pyramid Network decoder to produce the feature maps $\mathbf{F}_m$ at resolution 56×56.

Gaze subnetwork $\mathcal{G}$. The Human-Centric branch, on the other hand, uses a ResNet-18 backbone (11M parameters) [13] equipped with a custom 2D gaze prediction head. This sub-network operates on the head crop of the target subject, resized to 224×224.

Prediction module. Throughout the prediction module (cf. Figure 2), the feature maps $\mathbf{F}_m$, $\mathbf{T}_m$, and $\mathbf{F}$ are maintained at a resolution of 56×56, except in the regression subnetwork, where $\mathbf{F}$ is first upsampled to 64×64 (which is also the size of the predicted gaze heatmap) before going through different convolution layers. The embedding vectors e_{gaze} (i.e. from the head crop encoder), e_m (within the attention module) and e_{scene} (within the In-Out prediction head) each have a size of 512.

4. Experiments

4.1. Experimental protocol

We use two datasets for our experiments, and rely on standard metrics and protocols for evaluation.

Datasets. The first dataset is the *GazeFollow* dataset [29]. It comprises a curated set of images from popular image datasets. It was initially annotated with the 2D gaze target location, eye location, and head bounding box for most people in the images. Later, Chong et al. [6] extended these labels with indications of whether the gaze target of a person is located inside or outside the image. Overall, the dataset contains annotations for around 130k people in 122k images. The test set consists of 4782 people (all looking inside the image) whose gaze was annotated by 10 annotators.

The second one is the *VideoAttentionTarget* dataset [7]. It contains 1331 video clips collected from 50 shows on YouTube. The annotation comprises head bounding boxes and either the 2D gaze target location or whether the attention target is outside the frame. The training and test sets contain respectively around 131k and 33k bounding boxes. In general, the VideoAttentionTarget dataset has higher resolution frames and more close up and front-facing views of people compared to GazeFollow.

Training Protocol. To generate the ground-truth gaze heatmaps $\mathbf{H}^{gt}$, we place a Gaussian centered on the ground-truth gaze point and use a $\sigma = 3$ standard dev. (at the 64×64 heatmap resolution). In terms of training, the backbone of the sub-network in the Human-Centric branch is pre-trained on the Gaze360 dataset [17] to predict a 3D gaze vector, while the backbones of the feature extractors in the Scene-Centric branch are pre-trained on ImageNet [31]. To train the multimodal models, we first train the individual modalities separately (see Sec. 4.2), and initialize their multimodal counterparts with the learned weights. Further, following the training protocol of [9], all experiments on VideoAttentionTarget use models initialized with weights learned from GazeFollow. For VideoAttentionTarget, we subsample frames during training and use every third frame to avoid redundancy. All models are trained end-to-end using the AdamW optimizer [20] with a learning rate of 1e-4 for our experiments on GazeFollow, and a learning rate of 1e-5 for our experiments on VideoAttentionTarget. The loss coefficients are set to 100 for λ_{gaze}, 0.1 for λ_{dir}, and 1 for λ_{io} and λ_{att}. We train for 35 epochs on GazeFollow, and 20 epochs (40 for the Multimodal model) on VideoAttentionTarget.

Performance Metrics. The typical metrics used to evaluate gaze target prediction are:

- **AUC**: The predicted gaze target heatmap is compared against a binarized version of the ground truth gaze target heatmap. This is used to plot a curve for the True Positive Rate vs the False Positive Rate. The AUC is the area under this curve, where 1 is perfect performance and 0.5 is random behavior.
- **Distance**: The predicted gaze location is compared against the ground truth location using an L2 distance. We assume that each image is of size 1×1 when computing the L2 distances. Hence, distance values range from 0 to $\sqrt{2}$, where a lower value is better. When multiple annotations are available for the gaze location (GazeFollow), we compute the minimum and average distances to aggregate across all ground-truth labels.
- **Average Precision** (AP) is used to evaluate classification performance for the in vs out of frame prediction.

The AP is computed across the entire test set, and the distance and AUC on the subset of images with a ground-truth gaze target located inside the frame.

4.2. Tested models

Individual modalities. To evaluate the strength of each modality, we tested the model by relying on a single modality as input. In this case, the scene module does not include the fusion mechanism, and the feature map $\mathbf{F}$ of that modality is used directly as input to the Prediction module.

Privacy approach. In this approach, the goal is to rely only on processed and anonymized input data. Concretely, the Human-Centric module takes as input the crop of the subject's head from the pose image rather than the input image. The head skeleton is treated as an RGB input, and no changes are made to the architecture or training protocol.

Late fusion. We also evaluate a late fusion scheme where the gaze cone image g_{cone} and the binary head mask h_{loc} are fused with the Scene-Centric stream later in the architecture. Specifically, the two images are first downsampled before being concatenated together with the feature map $\mathbf{F}_m$ of each modality separately. This is represented by the blue dashed line in Figure 2.

Skip connections. To evaluate the importance of retaining information from higher resolutions during the upsampling process in the decoder of the Scene-Centric branch, we train a model on the image alone, without skip connections.

Modality Dropout. To evaluate the importance of modality dropout during training, we train a multimodal model without modality dropout.

State-of-the-art. We compare the performance of our approach to different state-of-the art methods for this task. Specifically, we include models from Chong et al. [7], Lian et al. [18], Jin et al. [16] and Fang et al. [9]. Given that some works use a temporal variant of their model on VideoAttentionTarget, we include their static variant as well for the sake of fairness when comparing the results.

4.3. Results

Our results on the GazeFollow and VideoAttentionTarget datasets are summarized in Table 1 and Table 2.

Model	AUC↑	AvgDist↓	MinDist↓
Lian [18]	0.906	0.145	0.081
Chong [7]	0.921	0.137	0.077
Jin [16]	0.919	0.126	0.076
Fang [9]	0.922	0.124	0.067
Image	0.933	0.134	0.071
Depth	0.921	0.141	0.080
Pose	0.902	0.164	0.100
Multimodal	0.943	0.114	0.056
Depth-privacy	0.920	0.152	0.088
Pose-privacy	0.893	0.175	0.109
Multimodal-privacy	0.928	0.136	0.075
Image-NoSkip	0.932	0.133	0.073
Multimodal-NoMoDrop	0.941	0.115	0.057
Multimodal-Late	0.931	0.128	0.068

Table 1. Results for our models on the GazeFollow dataset. Best scores are given in red and second best scores are given in blue.

Model	AUC ↑	Dist ↓	AP ↑
Chong [7]-static	0.854	0.147	0.848
Chong [7]	0.860	0.134	0.853
Jin [16]	0.870	0.127	0.882
Fang [9]	0.905	0.108	0.896
Image	0.918	0.122	0.864
Depth	0.899	0.134	0.852
Pose	0.904	0.131	0.866
Multimodal	0.913	0.110	0.879
Depth-privacy	0.891	0.156	0.831
Pose-privacy	0.881	0.150	0.823
Multimodal-privacy	0.895	0.140	0.826
Image-NoSkip	0.906	0.133	0.857
Multimodal-NoMoDrop	0.905	0.118	0.874
Multimodal-Late	0.905	0.113	0.863

Table 2. Results on the VideoAttentionTarget dataset. Best scores are given in red and second best scores are given in blue.

Individual Modalities. As the image contains the most complete information, it logically has the best performance on both datasets. Surprisingly, the performance of the depth and pose modalities are not very far, esp. when compared to state-of-the-art methods. In general, pose might be more accurate than depth when gaze is on faces or hands, but much worse when gaze is on scene objects since pose images contain absolutely no scene information. We observe that according to all metrics, depth is better than pose on GazeFollow, while it is around the same on VideoAttentionTarget. This can be explained by the fact that VideoAttentionTarget has more gaze points on faces compared to GazeFollow.

Multiple Modalities. Our Multimodal model gives better results compared to using the image alone on both datasets. The improvements are mainly visible on the distance metrics, where error reductions of 10% to 21% (the MinDist on GazeFollow) are achieved, which might be explained by the improved localization accuracy due to the pose cue, as well as disambiguation from the depth cue.

Compared to the state-of-the-art, we can see that our approach performs better than existing methods on GazeFollow for all metrics (e.g. reduction of more than 10% on distance metrics compared to [9]). On VideoAttentionTarget, our results are in par with those of Fang *et al*. [9]. As Fang et al. [9] process the eye regions to infer the gaze direction (which we do not), and eyes are more clearly visible in the VideoAttentionTarget dataset (it contains higher resolution front facing faces), we hypothesize that adding such information in our model would further improve our results.

Qualitative examples and attention scores. Qualitative examples are provided in Figure 3, where the outputs of the single and multimodal models are displayed. The attention scores predicted in the multimodal case are also indicated on each image modality for the given example. As can be seen, these scores reflect somehow the reliability associated with each cue. More generally, on GazeFollow, the average attention scores are 0.41, 0.36, 0.23 for the image, depth and pose modality, and 0.37, 0.31 and 0.32 on VideoAttentionTarget, reflecting the higher importance (and accuracy) of pose in this dataset as described earlier.

Privacy setting. On GazeFollow, the depth and pose models have similar performance to their counterparts where the gaze direction is inferred from a head crop of the image rather than from the pose image. However, on VideoAttentionTarget, there is a drop of performance. We believe this is because in GazeFollow, eyes are less visible and there are many instances where the face is not visible (hence, the head pose is the only available cue). In contrast, VideoAttentionTarget contains higher resolution front-facing faces where eyes can bring additional gaze information. The multimodal version (using depth and pose) improves performance (esp. on distance metrics) compared to the single modalities, obtaining similar performance to the image only model on GazeFollow and comparable performance to most state-of-the-art baselines on VideoAttentionTarget.

Skip-connections. We believe that one reason for the superior performance of our models compared to the state-of-the-art is the use of skip connections during the upsampling process. Without this feature (NoSkip model in result tables), we observe that while the performance is unchanged on GazeFollow, there is a performance drop on VideoAttentionTarget, esp. on the distance metrics. This may be because the higher resolution images of VideoAttentionTarget contain higher details about the face (eyes, nose) or objects, and can thus benefit more from the skip connections to precisely regress the target gaze locations.

Late fusion. The late fusion of the gaze information (gaze

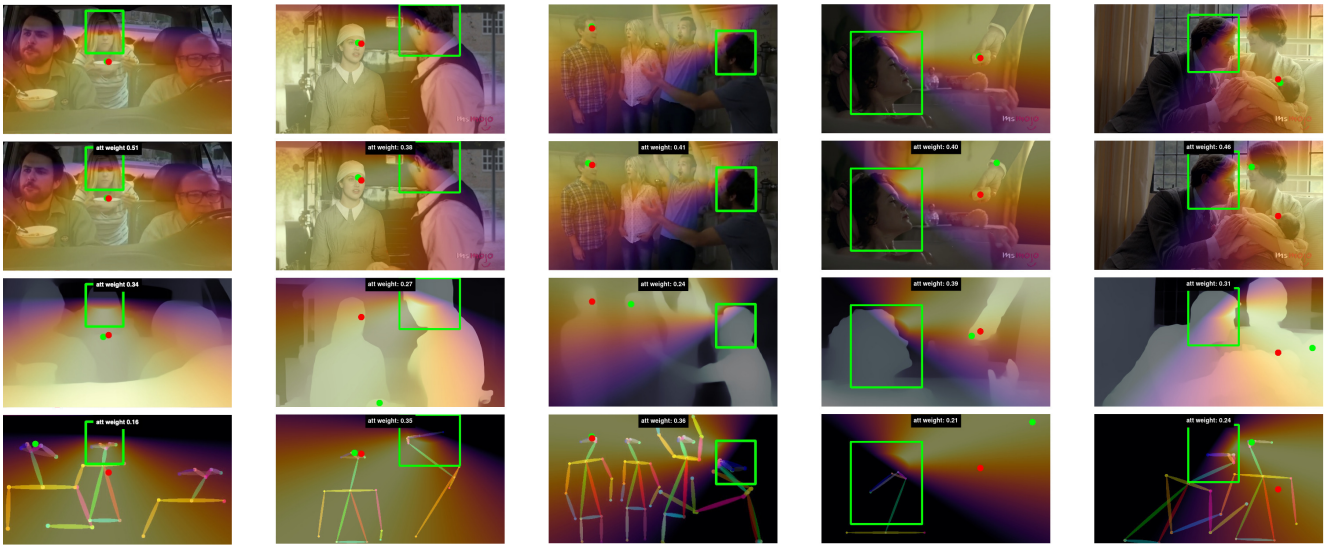

Figure 3. Qualitative results of our models (from top to bottom: multimodal, image, depth and pose). The image (or modality) is superimposed with the predicted gaze cone, the predicted gaze target (in **green**) and the ground truth target (in **red**). We observe that the attention scores reflect the reliability of the respective modalities for a particular sample (pose in 2nd and 3rd column; depth in 4th column), and that the fusion is able to ignore wrong information (pose in 1st and 4th columns; depth in 3rd column), and improve predictions of the image modality (4th and 5th column).

cone) with the feature maps rather than with the input images in our multimodal model improves performance compared to the models trained on a single modality (with early fusion), but has lower performance compared to the early fusion strategy in the multimodal case. We believe this is because introducing the person-specific gaze information early results in more capacity to identify the potential gaze target for that person. With the late fusion, the model has to identify potential gaze targets for all people in the scene (and at any place in the image).

Modality dropout. We analyze the importance of modality dropout during training. Without this feature, the multimodal model (NoMoDrop in the results tables) achieves a similar performance on GazeFollow, but obtains worse results on VideoAttentionTarget.

In the case of VideoAttentionTarget the average attention weights for image, depth and pose without modality dropout are 0.32, 0.29, 0.39, and the average weights with modality dropout are 0.37, 0.31, 0.32. Hence, modality dropout helps to learn a distribution of attention weights which better reflects the importance of the modalities. This may in turn help the model make better gaze target predictions.

5. Conclusion

In this paper, we proposed a modular multimodal architecture to explicitly leverage pose and depth information in order to improve the predicted gaze location and improve the state-of-the-art performance on two public benchmarks. We also investigated a late fusion scheme which allows us to first parse the scene in a person-agnostic manner, before introducing the subject's information. We showed that our model can also benefit privacy-sensitive applications in which personally identifiable information cannot be exposed. In this case, our model operates on head skeletons together with the pose and depth maps, achieving competitive performance.

Our architecture is modular and can naturally be extended to include other modalities, like optical flow (for videos), which we believe can further improve predictions. Alternatively, we can extend our model to inherently incorporate temporal information. Secondly, it is not clear at this point whether the depth cue is used as a way to verify the depth compatibility of the inferred gaze target with respect to the head position and gaze. Further study is needed to evaluate this. Finally, our current attention mechanism implies that one modality should be chosen to predict gaze. Conceptually, this formulation assumes the different modalities are equivalent, which is not necessarily the case. Thus, another future direction could investigate how to better fuse information across modalities.

Acknowledgement. This research has been supported by the ROSALIS project (Robot skills acquisition through active learning and social interaction strategies, grant agreement no. 30214) of the Swiss National Science Foundation (SNSF) as well as by the AI4Autism project (Digital Phenotyping of Autism Spectrum Disorders in children, grant agreement no. CRSII5_202235 / 1) of the the Sinergia interdisciplinary program of the SNSF.

References

[1] Sileye Ba and Jean-Marc Odobez. Recognizing visual focus of attention from head pose in natural meetings. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1), 2008. 2

[2] Sileye Ba and Jean-Marc Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *Transactions on Pattern Analysis and Machine Intelligence*, 33(1), 2011. 2

[3] Chongyang Bai, Srijan Kumar, Jure Leskovec, Miriam Metzger, Jay 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. 1

[4] Younes Belkada, Lorenzo Bertoni, Romain Caristan, Taylor Mordan, and Alexandre Alahi. Do pedestrians pay attention? eye contact detection in the wild, 2021. 3

[5] Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, page 1–10, New York, NY, USA, 2000. Association for Computing Machinery. 3

[6] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of the European conference on computer vision (ECCV)*, pages 383–398, 2018. 1, 3, 6

[7] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James 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. 2, 3, 4, 6, 7

[8] Lifeng Fan, Yixin Chen, Ping Wei, Wenguan Wang, and Song-Chun Zhu. Inferring shared attention in social scene videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6460–6468, 2018. 1, 3

[9] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao 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. 2, 3, 6, 7

[10] Sebastian Gorga and Kazuhiro Otsuka. Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction on - ICMI-MLMI '10*. ACM Press, 2010. 2

[11] Ralph Gross, Latanya Sweeney, Jeffrey Cohn, Fernando de la Torre, and Simon Baker. *Face De-identification*, pages 129–146. Springer London, London, 2009. 3

[12] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. Enhanced gaze following via object detection and human pose estimation. In *International Conference on Multimedia Modeling*, pages 502–513. Springer, 2020. 1, 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[14] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. We know where they are looking at from the rgb-d camera: Gaze following in 3d. *IEEE Transactions on Instrumentation and Measurement*, 2022. 3

[15] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pages 565–578. Springer, 2019. 3

[16] Tianlei Jin, Zheyuan Lin, Shiqiang Zhu, Wen Wang, and Shunda 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. 1, 3, 6, 7

[17] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 1, 6

[18] Dongze Lian, Zehao Yu, and Shenghua Gao. Believe it or not, we know what you are looking at! In *Asian Conference on Computer Vision*, pages 35–50. Springer, 2018. 1, 2, 3, 4, 6, 7

[19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 2, 4

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 6

[21] Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew 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. 3

[22] Benoit Masse, Sileye Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2711–2724, 2018. 2

[23] Zhixiong Nan, Jingjing Jiang, Xiaofeng Gao, Sanping Zhou, Weiliang Zuo, Ping Wei, and Nanning Zheng. Predicting task-driven attention via integrating bottom-up stimulus and top-down guidance. *IEEE Transactions on Image Processing*, 30:8293–8305, 2021. 1, 3

[24] Carman Neustaedter, Saul Greenberg, and Michael Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Trans. Comput.-Hum. Interact.*, 13(1):1–36, mar 2006. 3

[25] E.M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE Transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005. 3

[26] Kazuhiro Otsuka, Keisuke Kasuga, and Martina 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. 1, 2

[27] Kazuhiro Otsuka, Junji Yamato, Yoshinao Takemae, and Hiroshi Murase. Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In *2006 IEEE International Conference on Multimedia and Expo*, pages 949–952. IEEE, 2006. 2

[28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 5

[29] Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems (NIPS)*, 2015. * indicates equal contribution. 3, 6

[30] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1435–1443, 2017. 1, 2

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 6

[32] Samira Sheikhi and Jean-Marc 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. 1, 2

[33] Stephen V Shepherd. Following gaze: gaze-following behavior as a window into social cognition. *Frontiers in integrative neuroscience*, 4:5, 2010. 2

[34] Rémy Siegfried and Jean-Marc 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. 2

[35] Rainer Stiefelhagen, Michael Finke, Jie Yang, and Alex Waibel. From gaze to focus of attention. In *International Conference on Advances in Visual Information Systems*, pages 765–772. Springer, 1999. 2

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 5

[37] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard 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. 1, 3

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[39] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 5

[40] Yu Yu, Gang Liu, and Jean-Marc Odobez. Deep multitask gaze estimation with a constrained landmark-gaze model. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018. 1

[41] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution vision transformer for dense predict. *Advances in Neural Information Processing Systems*, 34, 2021. 5

[42] Zeynep Yücel, Albert Ali Salah, Çetin Meriçli, Tekin Meriçli, Roberto Valenti, and Theo Gevers. Joint attention by gaze interpolation and saliency. *IEEE Transactions on cybernetics*, 43(3):829–842, 2013. 2

[43] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, 128(5):1076–1100, 2020. 1, 3

[44] Lonnie Zwaigenbaum, Jessica A Brian, and Angie Ip. Early detection for autism spectrum disorder in young children. *Paediatrics & Child Health*, 24(7):424–432, 2019. 2