2024 18th International Conference on Automatic Face and Gesture Recognition (FG)

Visual Saliency Guided Gaze Target Estimation with Limited Labels

Cheng Peng and Oya Celiktutan
Department of Engineering, King's College London, London, UK

***Abstract*— Current models of gaze target estimation can present excellent performance, but the success of these models relies on large-scale annotated datasets. In real-world applications, obtaining large amounts of labelled data is often impractical due to the high cost of annotation. Therefore, in this paper, we investigate a relatively unexplored problem and introduce a semi-supervised method for gaze target estimation, which uses a small number of labels without compromising performance. We achieve this by leveraging the visual saliency map, which has been widely used in previous gaze target estimation studies. More explicitly, unlike previous studies, we build a multi-task model which can learn visual saliency and gaze target simultaneously. To train this model, in the lack of real labels, we propose a method to generate pseudo labels by combining the state-of-the-art approaches for visual saliency estimation, object detection, and head pose estimation. First, we train the multi-task model with the pseudo labels. Then, to compensate for the information loss due to the lack of reliable annotation, we fine-tune the network using a small number of real labels. We validate the performance of our model by creating a set of baseline models for comparison on two publicly available datasets, namely, GazeFollow and VideoAttention. The experimental results show that our method achieves the best performance in semi-supervised settings, as well as a competitive performance as compared to the existing fully supervised models. The code of the proposed method is available at https://github.com/PengC98/Weakly-supervised-gaze-target-estimation**

I. INTRODUCTION

Gaze behaviour is an important social signal in human-human communication because of its rich non-verbal information [8]. Therefore, how to automatically infer the visual attention of people in third-person images has attracted significant attention from the multidisciplinary research community. Automatic gaze analysis plays a crucial role in psychological research [7], [26], teaching quality assessment [41], [29], driving assistance [17], [39], and human movement analysis [10], [21], [33].

In recent years, many methods have been proposed by researchers to solve the problem of gaze target estimation, such as [30], [4], [5], [16], [25], [12], [22]. These methods show excellent performance on publicly available datasets, such as GazeFollow [30] and VideoAttention [5]. However, the success of these methods relies on large amounts of labelled data. In reality, acquiring large-scale data and annotation is often very expensive and labour-intensive. Therefore, it is increasingly needed to develop models that can understand images with minimal supervision in gaze target estimation and beyond.

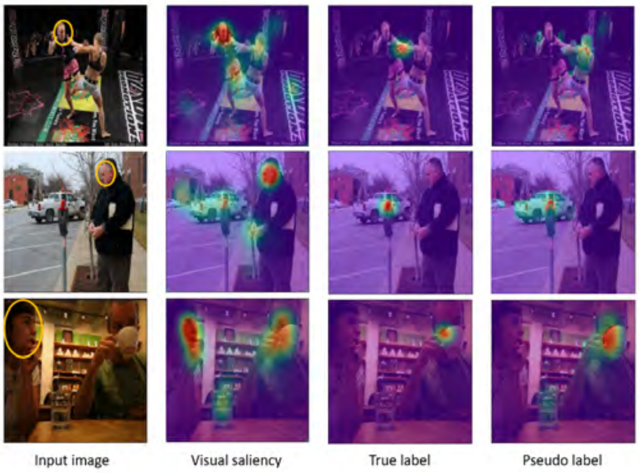

Fig. 1. Visualization of the visual saliency map, real label and pseudo label. In the first column, the yellow circle indicates the target person whose gaze is being predicted.

As shown in [30], [5], there is a synergy between visual saliency estimation and gaze target estimation tasks. More explicitly, in most cases, there is a relationship between the objects that people look at in a scene and the objects that are salient in it. To address the aforementioned gap in the literature, in this paper, we tap into this relationship and propose a semi-supervised method for gaze target estimation using visual saliency information.

To achieve this, we develop an approach to tackle two important challenges. First, it is a challenge to introduce visual saliency information into the gaze target estimation task. In order to make full use of the visual saliency information, unlike the methods proposed in [30], [4], we formalise it as a multi-task learning problem. Specifically, we build a model with a two-branch architecture. This model performs both visual saliency estimation and gaze target estimation given the same image as input. This structure forces the model to distil knowledge from the visual saliency prediction task and use it for the gaze target estimation task.

Second, to train the multi-task model in the lack of real labels, we propose a method for generating pseudo labels. As shown in Fig. 1, the visual saliency map contains not only information related to the gaze target but also a large amount of unrelated information. Therefore, the visual saliency map cannot be used as pseudo labels directly. To address this issue, inspired by [1], we scan the entire image to identify all potential gaze regions related to the target. In contrast to [1], who calculate front-most points from 3D spatial relationships between pixels, our approach aims to go beyond detecting the

The work of Cheng Peng has been supported by the King's China Scholarship Council (K-CSC) PhD Scholarship program.

979-8-3503-9494-8/24/$31.00 ©2024 IEEE

front-most points and semantically understanding the images. We start with semantic segmentation of the image, treating the instance segmentation masks as the possible gaze regions for the target person. Then, we integrate the chosen regions with visual saliency information to create pseudo labels. Our intuition is that the target person's attention should be correlated with the salient objects intersecting with their gaze direction [30]. These pseudo labels together with the visual saliency maps are used to train the multi-mask model, helping our model learn how to exploit and distinguish visual saliency information in the absence of real labels.

In order to show the performance of our method, we design three experiments. Firstly, we build three baseline models based on the current state-of-the-art models of gaze target estimation and compared them with our method under semi-supervised conditions. The experiment demonstrates that our model can show superior performance compared to baseline models. Secondly, we compared our models with current state-of-the-art fully supervised models. In the absence of labels, our model even achieves the performance of some fully supervised approaches. Finally, we also design ablation experiments to show the necessity of introducing visual saliency information and pseudo-labels.

Taken together, the main contribution of our paper can be summarized as follows:

- We present a semi-supervised learning framework for gaze target estimation. To the best of our knowledge, this is the first semi-supervised learning method applied to this domain.
- We propose a novel multi-task learning approach that can use visual saliency information to guide model inference about gaze information.
- We introduce a method for pseudo-label generation, which can be used to aid the model training.
- Our method exhibits the best performance in comparison with the baseline models. It even outperforms some fully supervised learning models, when a small number of real labels are used for training only.

II. Related Work

A. Gaze Target Estimation

The task of gaze target estimation was first proposed in [30], which aimed to enable the computer to recognize the object being observed by the target person, given an image and their head position. In the early years, the researchers made full use of the 2D image data to construct the model by extracting the person's face features and scene features separately and then combining these two features [30], [4], [5], [38], [31], [22]. Among them, Recasens et al. [30] and Chong et al. [4] treated this task as a one-hot patch classification. However, the scale of the patches is very coarse and can introduce errors in the model predictions. Then, in [22], Lian et al. transformed the gaze target estimation task into a heat map regression task. Meanwhile, instead of converting the head features into a latent space like in previous studies, they predicted the pose of the head image to generate a Gaze Direction Field, which can enhance the robustness of the network. In a recent study, Miao et al. [25] combined patch prediction and heat map regression to propose patch distribution prediction to improve the performance of the model on datasets with large variances in annotations. All the methods listed above can only estimate the gaze target of one person at a time. To address this, Jin et al. [16] proposed a model that can handle multiple people in the scene at the same time, and introduced a numerical regression method to reduce the quantification error caused by heat map prediction. After that, Tu et al. [37] proposed an end-to-end transformer-based model to predict the gaze targets and their head positions and postures for all the people in the scene.

There is another line of work focusing on gaze estimation in 3D using 2D information. Fang et al. [9] first estimated the depth information of the scene from 2D images, and then proposed a Dual Attention Module model combining the 2D field of view of the target person and the depth information of the scene. Bao et al. [1] also utilised this idea, but with the difference that after estimating the 3D point cloud using 2D images, they used the 3D gaze direction information to estimate the probability of the front-most point in the point cloud, which was being gazed at by the target person. Hu et al. [15] constructed a model based on a graph neural network to restore the positional relationship of each object in the scene using the estimated depth information. They inferred the probability of gazing at an object in the scene from these positional relationships.

However, a large amount of annotation data is necessary for the methods described above. In fact, fully annotating social interaction images with cluttered backgrounds is a very difficult task that requires massive resources. Therefore, the introduction of semi-supervised learning is essential to reduce the effort and cost of labelling a large number of images, hence expanding the practical applications.

B. Semi-supervised Learning

In reality, it is easier to collect data than to annotate them. Therefore, in order to make full use of the large amount of unlabelled data, semi-supervised learning has gained a lot of attention from researchers [3], [42], [20], [19], [35], [40], [2], [34]. In the field of gaze estimation, a number of semi-supervised and weakly supervised methods are also emerging. In [17], Kasahara et al. constructed a semi-supervised learning framework that can predict the target of a driver's gaze by exploiting the consistency of the driver's gaze direction with the salient information of the scene. Ghosh et al. [11] used off-the-shelf face image analysis models to create pseudo-labels for unlabelled faces based on multiple complementary auxiliary signals. These pseudo-labels were learned by a semi-supervised multi-task model with noise modelling. In contrast, Park et al. [28] introduced the concepts of representation learning and meta-learning, which allowed the model to learn a generalisable latent feature representation using a small number of samples. Kothar et al. [18] constructed a weakly supervised learning

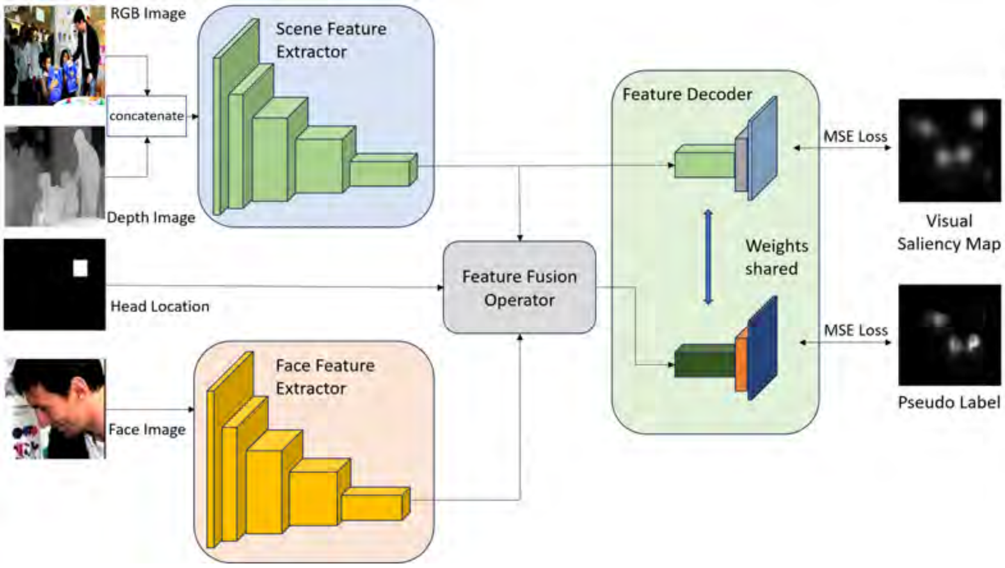

Fig. 2. Structure of the gaze target estimator

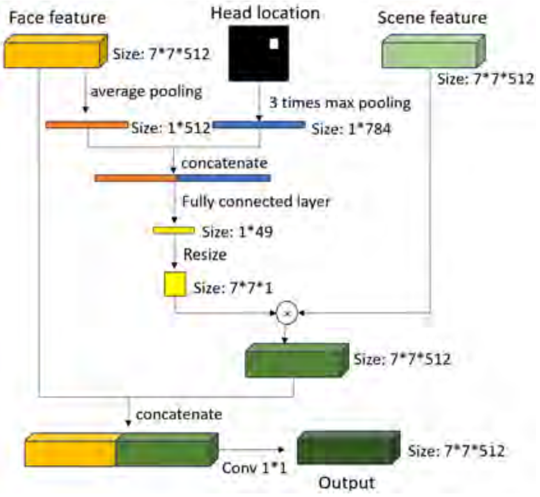

Fig. 3. Structure of feature fusion operator

model using geometric constraints between gaze directions when people look at each other in an interactive scene.

However, the task of gaze target estimation places higher demands on the model's ability to understand the scene. The auxiliary signals applied in the methods described above can hardly help the model understand the scene information in the absence of labels. Therefore, in our work, we develop a multi-task learning model to enhance scene understanding through a visual saliency estimation task. Concurrently, we design a pseudo-label generation mechanism, which allows the model to obtain a full comprehension of the scene with a small number of real labels.

III. Methodology

Our methodology consists of two main stages. Firstly, in order to enable our model to obtain information from the visual saliency estimation task, we build a novel gaze target estimator with a multi-task learning architecture. This model uses the same feature decoder to decode different features from the same image and output both visual saliency map and gaze target estimation map. Secondly, we propose a pseudo-label generation method to train this architecture without real labels. Specifically, we strategically identify areas on the visual saliency map corresponding to potential gaze locations based on the head pose of the target person. By training our multi-task model using pseudo-labels generated in this way, we can further encourage the gaze target estimation component of the model to learn relevant information from the visual saliency estimation task.

A. *Gaze Target Estimator*

Fig. 2 shows the whole structure of our model and its inputs and outputs. This model takes four inputs: the RGB image, the depth image obtained by [36], a crop of the target person's face, and the location of the face which is a binarized image of the size same as the RGB image, with 1 indicating the face position and 0 indicating the rest of the area. Based on these inputs, the model estimates the visual saliency of all objects in the scene and the likelihood of being gazed at by the target person. As shown in Fig. 2, the model consists of two fully convolutional feature extractors, a feature fusion operator and a decoder. The feature extractors extract features from the scene and the target person's face separately. In [4], it has been shown that humans naturally follow where others are looking by observing their head posture. In analogy with this, we design two feature extractors dedicated to the scene and the target person's face individually. The feature fusion operator combines the features from the scene and face branches, enabling the model to reason about the scene information according to the face information. The structure of the feature fusion operator is given in Fig. 3. It takes the face feature, scene feature, and head location map as input. These inputs are passed through

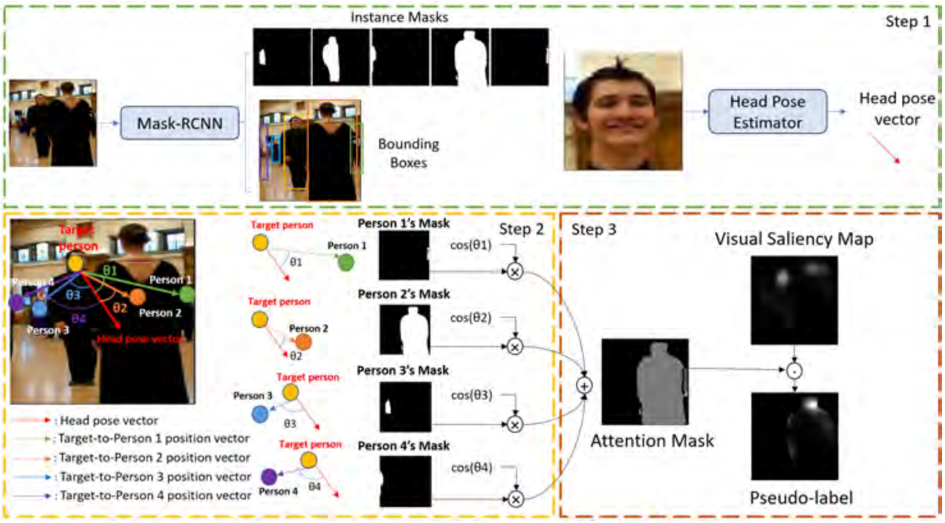

Fig. 4. Pseudo-label generation. There are three steps to build the pseudo-label: 1) instance segmentation and head pose estimation; 2) angular similarity calculation; and 3) attention mask generation. We segment the objects in the scene using Mask-RCNN to obtain the pixel coordinates of each object and estimate the head pose of the target person by an on-the-shelf head pose estimator [32]. Then, we calculate the likelihood (defined as Cosine Similarity) of each object being looked at by the target person based on the target's head orientation. The attention mask is constructed by combining the calculated likelihoods with the instance-level segmentation masks. Finally, the attention mask and the visual saliency image are overlaid to obtain the pseudo-label.

one fully connected layer and two convolutional layers to produce a feature of the same size as the input scene feature. As for the decoder, it is a standard Feature Pyramid Network (FPN) [23], which is widely used in object detection tasks, and was first applied by [22] in the field of gaze target estimation. It up-samples the input features, synthesises the intermediate layer features from the scene feature extractor, and finally outputs the heatmap through a sigmoid activation function.

Training Process. As shown in Fig. 2, we begin with extracting scene features and face features using the scene and face feature extractors, respectively. Then, these features are sent to the feature fusion operator, which outputs a fused feature. Following this fusion step, both the fused feature and the original scene feature are fed into the feature decoder to generate the gaze heatmap and the visual saliency map simultaneously. To be more specific, the feature decoder adapts its decoding process based on the type of input features it receives. When provided with the scene feature as input, the feature decoder generates a visual saliency map, supervised by the output of a pre-trained method SalGAN [27]. In contrast, when given the fused feature as input, the decoder transforms it into a gaze heatmap, supervised by pseudo-labels as explained in the following section first. After pre-training with the pseudo-labels, a small amount of real labels is used to fine-tune the model. Importantly, this fine-tuning process occurs while the model continues to perform the visual saliency estimation task.

B. Pseudo Label Generation

In the lack of real labels, to train the gaze estimator proposed in Section III-A, we develop a pseudo-label generation method. Fig. 4 shows the three steps of building pseudo-labels.

In the first step, given an image with a target person whose gaze is to be estimated, our objective is to determine the number of objects within the scene, which could potentially attract the target person's attention. To do so, we identify and segment all objects and people in the scene using Mask-RCNN [13], generating N instance-level segmentation masks and bounding boxes for these instances, including the target person. Therefore, there are $N-1$ objects that might be gazed at by the target person. We can also calculate the position of each object from the output bounding boxes. The bounding box of an object i can be represented as $(x^i_{ul}, y^i_{ul}, x^i_{lr}, y^i_{lr}), i \in N-1$. The instance-level segmentation mask of an object is M_i. Then, we estimate the head pose of the target person using the method proposed in [32] and obtain a two-dimensional vector F_{hp} in pixel coordinates.

For the second step, we define a method to calculate the probability that each object in the scene is being gazed at by the target person. Determining whether an object attracts the target person's attention typically depends on the alignment with their gaze direction. The smaller the angle between the spatial position vector from the target person's head location to the estimated object and the vector of the target's gaze direction, the greater the probability that the object is being gazed at. Thus, we define the gazing probability P based on the cosine similarity between the two vectors. A smaller value of P means that the angle between the two vectors is larger. The direction vector from the object in the scene to the target person can be expressed as $F^i_r = c^i_o - c_p$, where c_o and c_p are the object and the target person's location in pixel coordinates. The location of the object c_o can be calculated as $c^i_o = [\frac{x^i_{ul}+x^i_{lr}}{2}, \frac{y^i_{ul}+y^i_{lr}}{2}], i \in N-1$. The location of the target person c_p is represented by the target person's head location. Then we can use the normalised person-to-object vector F^i_r and the target's head pose vector F_{hp} to calculate the gazing probability P_i for each object as follows:

$$P_i = \frac{1 + \frac{F_{hp} \cdot F_r^i}{||F_{hp}|| \times ||F_r^i||}}{2}, i \in N-1 \tag{1}$$

After obtaining the gazing probabilities, we can generate an attention mask A_m from the instance-level segmentation masks $M_i, i \in N-1$ for each object that the target person is likely to gaze, as shown in Fig. 4 (refer to Step 2). In other words, the attention mask can be interpreted as the composition of individual instance-level segmentation masks, weighted by the gazing probabilities, and it can be calculated as follows:

$$A_m = \begin{cases} \frac{1}{N-1} \sum_{i=1}^{N-1} \left(M_i \cdot ReLU(P_i - \frac{1}{N-1} \sum_{j=1}^{N-1} P_j) \right), & \text{if } N > 2 \\ \frac{1}{N-1} \sum_{i=1}^{N-1} (M_i \cdot P_i), & \text{if } N = 2 \\ 1 - M_1, & \text{if } N = 1 \end{cases} \tag{2}$$

As shown in the equation above, we set three cases. When $N = 1$, only the target person appears in the image. In this case, any region within the image, except for the target person's location, may serve as the gaze target. In the case where $N = 2$, the image contains only one additional object and the target person. In this case, this additional object may become the gaze target. For the case $N > 2$, we calculate the average gazing probabilities across all objects in the image. Only objects with gazing probabilities surpassing the calculated mean value are selected as potential gaze targets.

Finally, in step three, we know that the probability of an object being gazed at is also likely to be related to whether the object is salient in the scene. Due to the uncertainty in the estimation of the head pose, the head orientation can not fully represent the gaze direction. So some objects may be more attractive and more likely to be gazed at by the target person even though they are farther away from the gaze direction because they are very salient in the scene. Therefore, in generating pseudo-labels, we overlaid the attention mask on the visual saliency map obtained from SalGAN [27], as shown in Fig. 4 (refer to step 3). The resulting pseudo-label l_p can be represented as $l_p = V_s \cdot A_m$, where V_s is the visual saliency map. In addition, this pseudo-label generation method is suitable for our multi-task model. These pseudo-labels, intercepted and weighted from the visual saliency maps, can help our model gain the ability to extract information from the visual saliency task better in the absence of real labels.

IV. Experimental Results

A. Implementation Details

In this section, we describe the implementation details of the proposed model. The inputs to the model are the RGB image, the depth image, the face image of the target person who needs to be estimated and a binary image representing the target person's head location. These four images are all resized to 224×224. This model consists of a scene feature extractor, a face feature extractor, a feature fusion operator and a decoder. Resnet34 [14] is used as the backbone network for extracting the scene feature and the face feature. Specifically, for the scene, the input layer of Resnet34 is extended to 4 channels to accommodate the input of RGB images and depth images. For the face image, the convolution layer of Resnet34 is retained completely. Therefore, the output feature size of both convolutional paths is $7 \times 7 \times 512$. In the feature fusion operator, we perform three times Maximum pooling for the head location map and compress it into a vector of size 1×784. Then, we input the face features into the Average pooling layer to make a 1×512 vector and concatenate it with the head location vector, resulting in a vector of size 1×1296. This vector will be sent into a fully connected layer to get a self-attention weight which will weigh the scene feature. After that, we concatenate the weighted scene feature and the face feature. Finally, the concatenated feature will be fed into two convolutional layers with kernel size of 1×1 to compress the feature's size from $7 \times 7 \times 1024$ to $7 \times 7 \times 512$. The loss function we used in our training is the Mean Squared Error (MSE) loss.

B. Dataset

In this paper, we test our model on two of the most common publicly available datasets, GazeFollow [30] and VideoAttention [5].

The GazeFollow dataset consists of several publicly available image datasets, including ImageNet [6], MS COCO [24], and others. The Gazefollow dataset contains 122,143 images and 130,339 target characters, of which 4782 target characters were split as the test set and the rest were used as the training set. To ensure annotation quality, each sample in the test set contained 10 annotations from different people. Recently, [36] extended it with depth information for all samples.

The VideoAttention dataset was created by selecting 50 films from a large number of easily accessible film and television drama sources and extracting some of the clips. The duration of these clips varies from 1 to 80 seconds and contains a total of 164,541 frames. Each frame is annotated in detail with bounding boxes and gaze target points. 31,978 frames of short clips from 10 films and TV dramas were selected as the test set.

C. Evaluation Metrics

In order to ensure a fair comparison, we use three evaluation metrics by following [30]. **AUC**: We use the area under curve (AUC) criteria to assess the confidence level of the predicted heat map. **Dist.**: We calculate the Euclidean distance between the predicted gaze point and the real label. **Ang.**: We calculate the angular error between the predicted gaze direction and the ground truth gaze vector.

D. Experimental Setup

We use the PyTorch framework for implementing the proposed method. All the training sessions are run on an RTX3090 Graphic card. The hyper-parameters we set are as follows: batch size(20), learning rate(0.0001). Adaptive

TABLE I

QUANTITATIVE RESULTS OF COMPARISON WITH THE BASELINE MODELS

Dataset	GazeFollow												Videoattention							
Method	10%			15%			25%			50%			10%		15%		25%		50%	
	AUC	Dist.	Ang.	AUC	Dist.	Ang.	AUC	Dist.	Ang.	AUC	Dist.	Ang.	AUC	Dist.	AUC	Dist.	AUC	Dist.	AUC	Dist.
MT-Video	0.810	0.257	38.1	0.822	0.254	37.3	0.852	0.238	34.3	0.885	0.187	28.3	0.750	0.233	0.759	0.218	0.763	**0.216**	0.800	0.194
MT-Lian	0.834	0.242	32.6	0.871	0.211	**28.9**	0.877	0.210	29.8	0.890	0.176	25.4	0.808	0.231	0.815	**0.216**	0.825	0.222	0.840	0.207
MT-Ours	0.845	0.237	34.8	0.870	0.210	30.5	0.883	0.196	29.0	0.891	0.183	26.2	0.765	**0.218**	0.785	0.222	0.792	0.220	0.806	0.211
Ours w/o depth	0.867	0.228	32.5	0.879	0.200	29.7	0.887	0.192	28.7	0.896	0.179	25.6	0.813	0.228	0.820	0.226	0.831	0.219	0.842	0.194
Ours	**0.871**	**0.217**	**30.1**	**0.883**	**0.200**	29.2	**0.890**	**0.189**	**28.6**	**0.898**	**0.176**	**25.1**	**0.818**	0.220	**0.825**	0.229	**0.834**	0.217	**0.846**	**0.191**

moment estimation (Adam) optimizer is used to train our network.

We train our gaze estimator model in a semi-supervised manner with a small amount of labelled images. For the GazeFollow dataset, we randomly select 10%, 15%, 25%, and 50% of the labelled training samples, while the rest of the data is used as unlabelled data. For the VideoAttention dataset, we divide labelled and unlabelled data using a different approach. There are multiple clips from 40 different movies in the training set of the Videoattention dataset, and we randomly select all the clips from 4, 6, 10, and 20 of these movies as labelled data, which respectively corresponds to 10%, 15%, 25%, and 50% of the labelled training samples. We train our model directly on the VideoAttention dataset without pre-train on the GazeFollow dataset.

It should be noted that, as shown in Fig 2, our method needs the location of the head. To train our method we use the ground truth values for the head location and add perturbations to ensure robustness against potential head localization errors.

Since this is the first time that a semi-supervised learning approach has been used in the field of gaze target estimation, we build three baseline models for comparison to validate our model. We use the Mean-Teacher method [35], which is commonly used in the field of semi-supervised learning, to make some modifications to the current state-of-the-art models in gaze target estimation and train them in a semi-supervised way. The Mean-Teacher model learns to make predictions that are consistent with its past checkpoints (teacher model). This helps the model to generalise better and prevents the model from making unstable predictions in the face of unlabelled data. At the same time, compared to other semi-supervised frameworks, the Mean-Teacher method does not need to change the output form of the original model and does not need to introduce more hyper-parameters. Given a method that we wish to apply the Mean-Teacher framework, we only need to build another model identical to the original model, namely the teacher model. Then, the original model, or the student model, can be trained by the original loss function and optimizer. As for the teacher model, its parameters are updated by the exponential moving average (EMA) of the student model. In addition, the MSE loss needs to be added to the original loss function to ensure consistency between the student model and teacher model outputs. Since most of the advanced gaze target estimation methods' source code is not available, we select two of them and produce the following baselines. **MT-Video**: We use the Mean-Teacher method to retrain the model in [5]. In training, the original model hyper-parameters and model initialisation methods are retained. However, the MSE loss was added to the original loss function to ensure the consistency of the heatmap output between the teacher model and the student model. **MT-Lian**: We use the Mean-Teacher method to retrain the model in [22]. In training, the original model structure and loss function were not changed, only the MES loss function was added to maintain the consistency of the heatmap and gaze direction between the outputs of the teacher model and the student model. Besides this, the hyper-parameters are set as follows: batch size(20) and learning rate(0.00025). **MT-Ours**: In order to make a full range of comparisons, we make changes based on the Mean-Teacher approach using the structure of our proposed model. The original model hyper-parameters and model initialisation methods are retained. An MSE loss was added to our initial loss function to ensure consistency between the visual saliency heatmaps and the gaze target heatmaps output by the teacher model and the student model. However, in training this baseline model, we do not pre-train it using our proposed pseudo-label. **Ours w/o depth**: Since neither Lian [22] nor Video [5] introduces depth information, we trained our model without depth information as input for fair comparison.

E. Comparison with Baselines

Table I demonstrates the quantitative results as the comparison with the baseline models. The evaluation results on GazeFollow are shown on the left side of the table. We can see that our model can always achieve the best performance on the AUC metrics and Dist. metrics on the GazeFollow dataset. When trained with 10% real labels, our method improves the gaze estimation by a margin of 4.4% and 10.3% in terms of the AUC metric and the Dist metric as compared to the best performing of the MT-Video and MT-Lian. Our method also improves by a margin of 3.1% and 8.4% in terms of AUC metric and Dist. metric as compared to MT-Ours. Although the improvement in the performance of our model is reduced as compared with the three baseline models when trained with 15% and 25% real labels, it still improves

TABLE II

EVALUATION OF OUR MODEL AND OTHER FULLY SUPERVISED MODELS

Method	GazeFollow			Videoattention		Params.
	AUC	Dist.	Ang.	AUC	Dist.	
SVM + one grid [30]	0.758	0.276	43.0	-	-	-
SVM + shift grid [30]	0.788	0.268	40.0	-	-	-
Recasens [30]	0.878	0.190	24.0	-	-	-
Chong [4]	0.896	0.187	-	0.830	0.193	-
Lian [22]	0.906	0.145	17.6	0.867	0.168	52M
Video [5]	0.921	0.137	-	0.854	0.147	61M
Fang [9]	0.922	0.124	14.9	0.878	0.124	66M
Tonini [36]	0.927	0.141	-	0.940	0.129	92M
Bao [1]	0.928	0.126	15.3	0.885	0.120	63M
Miao [25]	0.934	0.123	-	0.912	0.109	61M
Ours - fully supervised	0.914	0.157	21.7	0.898	0.140	44M
Ours-resnet50 – fully supervised	0.908	0.160	22.9	0.897	0.141	60M

the AUC metric by 1.3% and 1.4%. The evaluation result on VideoAttention is shown on the right side of the table. Our model always achieves the best performance in terms of the AUC metric and improves by a margin of 1.2%, 1.2%, 1.1% and 0.7% as compared with the baselines, when trained with 10%, 15%, 25% and 50% of real labels. From the result shown in Table I, we can see that our method can achieve the best results 16 times out of 20 as compared with other baselines. Although our results are not the best on the Dist metrics compared to the other baseline on the VideoAttention dataset when we use 10% to 25% true labels, our results are pretty close to the best. And the Dist. metrics may be meaningless to some extent [1]. Additionally, because MT-Video and MT-Lian do not take depth information as input, to make a fair comparison, we evaluate our model without depth information. From the result, we can see that even without depth information, our model can achieve good performance. Thus, the newly added visual saliency information can already bring semantic understanding to our model without depth information.

F. Comparison with Fully Supervised Methods

While *our primary focus in this paper is on addressing gaze estimation under weakly-supervised conditions*, we conducted fully-supervised training for our proposed model to show its performance comprehensively. We present the results of comparison with current state-of-the-art fully supervised methods in Table II. It should be noted that on the VideoAttention dataset, we follow the training settings of the other models to ensure a fair comparison. We trained our model on the GazeFollow dataset until convergence, and then we fine-tuned our model on the VideoAttention dataset.

The quantitative results in Table II show that although we can not surpass some SOTA methods, our method remains competitively positioned in comparison with other methods proposed so far. Especially, when considered alongside with Table I, our model under weakly supervised conditions surpasses some fully supervised methods like Recasens [30] and Chong [4] with respect to all three metrics.

Compared with Video[5], Fang [9], Tonini [36], Bao [1] and Miao [25] which outperform our method under the fully supervised condition, our model has the least number of parameters. In contrast to our methods, most of the methods proposed so far use Resnet50 rather than Resnet34 as the backbone. Because of that, they also use a large number of parameters in constructing the attention mechanism to build the relationship between the scene image and the face image. Consequently, benefiting from the backbone with stronger generalisation capabilities and a more sophisticated attention mechanism, these methods can achieve performance that is better than our model on fully supervised training. However, it is difficult for such models with a large number of parameters to achieve good performance under semi-supervised training settings (i.e., using a small number of annotated samples only). As shown in Table I, the model with more parameters such as Video [5] is performing worse as compared to Lian [22] and our model when we use the same Mean-Teacher based structure and the same number of labelled samples for training.

TABLE III

THE RESULTS OF ABLATION STUDIES

Method	AUC	Dist	Ang
w/o.visual saliency prediction branch	0.840	0.245	33.5
w/o.pseduo-label training	0.856	0.228	31.8
Ours (10%)	0.871	0.217	30.1

We further investigated the performance of our approach using Resnet50 as the backbone, but it can be seen from Table II that there is a slight deterioration in the performance. We conjecture that this might be due to the overfitting problem in visual saliency estimation and thus our model struggles to transfer the knowledge from visual saliency estimation to gaze target estimation. our model cannot effectively fine-tune the pre-trained model with limited labels when a deeper model is used.

Besides the number of parameters, some of the models such as Lian [22], Fang [9] and Bao [1] additionally incorporate supervision from manually annotated gaze direction features, further improving the performance in terms of the Dist. and Ang. metrics. Although our model's performance with respect to the AUC metric is better than Lian [22], due to the lack of gaze direction information, our model performs slightly worse than Lian [22] with respect to the Dist. and Ang. metrics on the GazeFollow dataset, when we perform fully supervised training. However, estimating gaze direction is a problem per se, and solving it with a limited number of annotated samples is not trivial. In contrast, our approach relies more on scene reasoning, enabling competitive performance on the VideoAttention dataset after pre-training on the GazeFollow dataset.

G. Ablation Studies

In this section, we analyse the impact of each component on the model. In order to make a fair comparison, we design the following baselines and use 10% of the real labels for training.

- **Without visual saliency prediction branch**: We remove the branch of the visual saliency map prediction and let the network predict the gaze target directly.

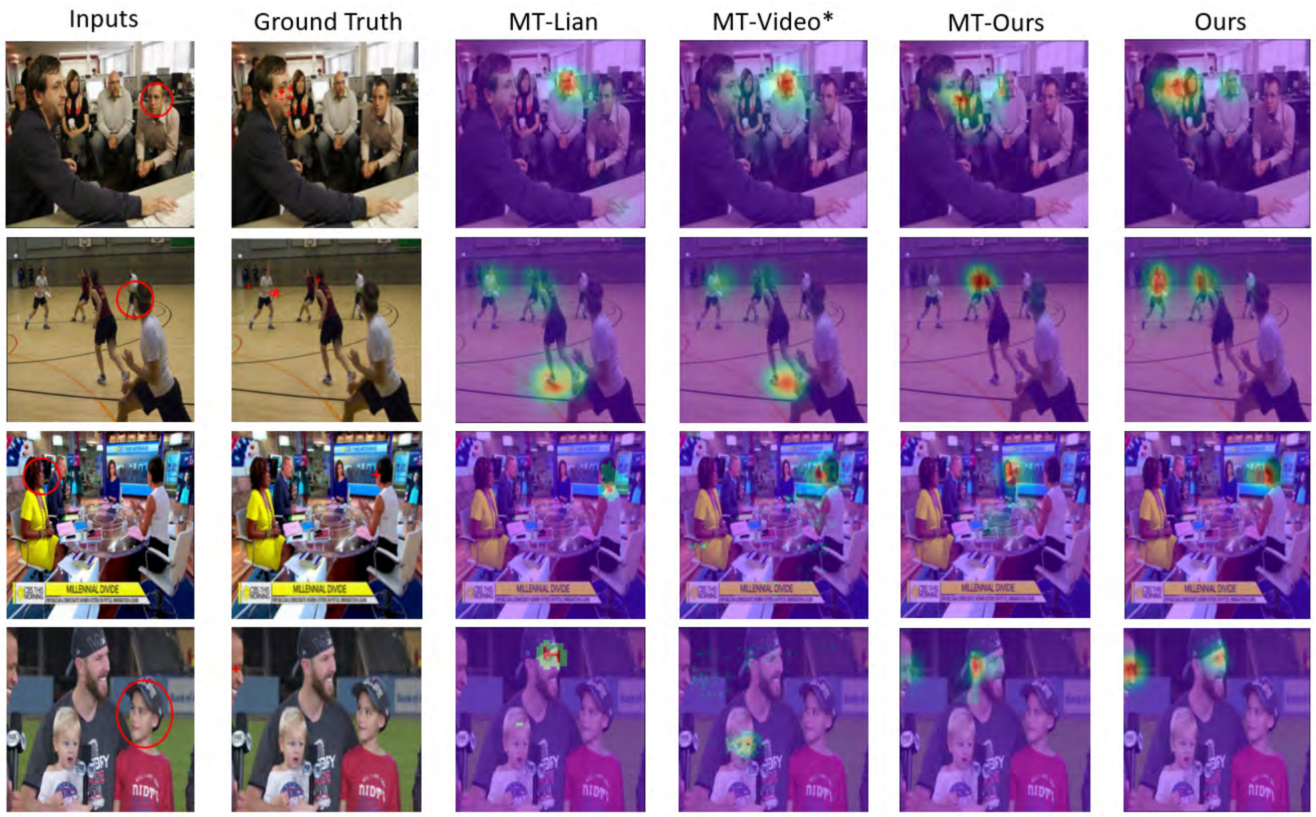

Fig. 5. Visualizations of the output of baseline models and our model (all trained by 10% of labels). The first two rows are on GazeFollow [30] and the last two rows are on VideoAttention [5]. In the first column, the red circle indicates the target person. The ground truth gaze target points are shown in the second column.

Specifically, the feature decoder does not take the scene feature as input directly to do the visual saliency estimation in this baseline. The model is trained by 10% of real labels directly without any pre-training steps.

- **Without pseudo-label training**: In this baseline, all the model components are retained. We only remove the step of pre-training the network using pseudo-labels. We train this model under the supervision of the pre-trained visual saliency estimator and 10% of the real labels.

Table III shows the results of the ablation studies. From the results, we can see that adding the branch of visual saliency estimation can effectively improve the results of gaze target estimation with limited labels. Pre-training the model using pseudo-labels can enhance the model's ability to understand semantic information better.

H. Visualization of prediction results

We visualize the outputs from the baseline model we produced and our model in Fig. 5. In line with the quantitative results, our model is more accurate as compared with these baseline models under limited supervision.

V. Conclusion

We propose a novel gaze target estimation model that can be trained using a small number of labels. To achieve this, we first construct a two-branch structure that allows the model to perform both the visual saliency estimation task and gaze target estimation task. This structure forces the model to utilise visual saliency information to do semantic understanding in the absence of label information. Subsequently, we construct pseudo-labels using the target's head pose and objects' location combined with the visual saliency map and pre-trained the model using the pseudo-labels to enhance the model's ability to utilise visual saliency features. From the results, our model shows an advantage in comparison with the baselines we constructed based on commonly used semi-supervised learning frameworks. At the same time, our model's performance can approach or even exceed the performance of some fully supervised deep-learning models.

References

[1] J. Bao, B. Liu, and J. Yu. Escnet: Gaze target detection with the understanding of 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14126–14135, 2022.

[2] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[4] E. Chong, N. Ruiz, Y. Wang, Y. Zhang, A. Rozga, and J. 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.

[5] 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.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] H. Duan, X. Min, Y. Fang, L. Fan, X. Yang, and G. Zhai. Visual attention analysis and prediction on human faces for children with autism spectrum disorder. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(3s):1–23, 2019.

[8] N. J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000.

[9] 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*, pages 11390–11399, 2021.

[10] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 314–327. Springer, 2012.

[11] S. Ghosh, M. Hayat, A. Dhall, and J. Knibbe. Mtgls: Multi-task gaze estimation with limited supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3223–3234, 2022.

[12] 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.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] 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.

[16] 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.

[17] I. Kasahara, S. Stent, and H. S. Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*, pages 126–142. Springer, 2022.

[18] R. Kothari, S. De Mello, U. Iqbal, W. Byeon, S. Park, and J. Kautz. Weakly-supervised physically unconstrained gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9980–9989, 2021.

[19] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

[20] D.-H. Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[21] Y. Li, M. Liu, and J. M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018.

[22] 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.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[25] 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.

[26] J. S. Oliveira, F. O. Franco, M. C. Revers, A. F. Silva, J. Portolese, H. Brentani, A. Machado-Lima, and F. L. Nunes. Computer-aided autism diagnosis based on visual attention models using eye tracking. *Scientific reports*, 11(1):10131, 2021.

[27] J. Pan, E. Sayrol, X. G.-i. Nieto, C. C. Ferrer, J. Torres, K. McGuinness, and N. E. OConnor. Salgan: Visual saliency prediction with adversarial networks. In *CVPR scene understanding workshop (SUNw)*, 2017.

[28] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz. Few-shot adaptive gaze estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9368–9377, 2019.

[29] N. S. Rebello, M. H. Nguyen, Y. Wang, T. Zu, J. Hutson, and L. C. Loschky. Machine learning predicts responses to conceptual tasks using eye movements. In *Physics Education Research Conference 2018, PER Conference, Washington, DC*, volume 10, 2018.

[30] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015.

[31] 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.

[32] N. Ruiz, E. Chong, and J. M. Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018.

[33] Y. Shen, B. Ni, Z. Li, and N. Zhuang. Egocentric activity prediction via event modulated attention. In *Proceedings of the European conference on computer vision (ECCV)*, pages 197–212, 2018.

[34] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[35] A. Tarvainen and H. Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[36] F. Tonini, C. Beyan, and E. Ricci. Multimodal across domains gaze target detection. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 420–431, 2022.

[37] 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.

[38] B. Wang, T. Hu, B. Li, X. Chen, and Z. Zhang. Gatector: A unified framework for gaze object prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19588–19597, 2022.

[39] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney. Predicting driver attention in critical situations. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V 14*, pages 658–674. Springer, 2019.

[40] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

[41] Q. Zhou, W. Suraworachet, O. Celiktutan, and M. Cukurova. What does shared understanding in students' face-to-face collaborative learning gaze behaviours "look like"? In *International Conference on Artificial Intelligence in Education*, pages 588–593. Springer, 2022.

[42] Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541, 2005.