Where are *they* looking in the 3D space?

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

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

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

Abstract

We propose a novel depth-aware joint attention target estimation framework that estimates the attention target in 3D space. Our goal is to mimic human's ability to understand where each person is looking in their proximity. In this work, we tackle the previously unexplored problem of utilising a depth prior along with a 3D joint FOV probability map to estimate the joint attention target of people in the scene. We leverage the insight that besides the 2D image content, strong gaze-related constraints exist in the depth order of the scene and different subject-specific attributes. Extensive experiments show that our method outperforms favourably against existing joint attention target estimation methods on the VideoCoAtt benchmark dataset. Despite the proposed framework being designed for joint attention target estimation, we show that it outperforms single attention target estimation methods on both the GazeFollow image and the VideoAttentionTarget video benchmark datasets.

1. Introduction

We live in a multi-dimensional world and experience our environment through our senses. Vision is one of the most dominant ways we experience the world. Our sense of orientation within a given context, location, and place is determined by an animated 3D model of the world our brains construct from the cues around us. Image-based gaze target estimation aims to infer what the subjects in the image scene are looking at from a single RGB image. Human gaze following and gaze target estimation in the wild are fundamental for visual navigation. Furthermore, this information is important to evaluate intentions and predict human behaviours in various social contexts [7]. For these reasons, gaze analysis has widely been used in neurophysiology studies [6, 21], relevant saliency prediction [5, 20] and social awareness tracking [4, 18, 19]. Humans are naturally good at understanding the actions of others and estimating where they are looking by leveraging prior knowledge. They can infer the pose and the person's orientation and reconstruct the 3D image scene based on a single view.

Figure 1. **Attention target estimation example use case visualisation.** The driver performs gaze following of the pedestrians to infer their intention and prevent a potential collision.

By looking at the image, they can understand the person's actions and infer their intentions. They can even guess what it would look like from another viewpoint. We can do this because all the previously seen objects and scenes have enabled us to build prior knowledge and create mental models of object appearance.

Figure 1 shows an example of an everyday scenario where the human gaze following is critical to safe driving. In this instance, the vehicle's driver approaches a junction where two pedestrians are about to cross the road. A crucial part of safe driving is adequate situational awareness of the driver. This includes predicting the actions of other road users, such as pedestrians and cyclists within their proximity and on their path. Gestures such as pedestrians looking around for traffic could indicate their intention to cross the road and potentially the driver's path. Furthermore, this enables the driver to determine whether the pedestrian has spotted them, which would help to prevent a potential collision. To do so, the driver observes the pedestrians from a third-person view and estimates their individual and joint gaze direction and target.

From the neurocognitive perspective, gaze perception is performed by humans to discriminate the gaze direction of others [27] as part of various social interactions, such as gaze following. This action is a vital part of the cognitive functions that allow people to learn via observation [10]. Once they successfully perceive the gaze direction of others, they can utilise this knowledge through their social cognition system in various ways [25], for example, to engage in joint attention with the observed person. By definition, joint attention happens when a gaze leader looks at a particular object which induces gaze followers to orient their attention to the same target. In computer vision, the task of joint attention target estimation is often referred to as shared attention in the literature [3, 8, 12]. While these terms are similar, they are subtly different from each other [7, 25]. In this work, we define joint and shared attention terms according to the neurocognitive perspective as in [7] and treat gaze perception and joint attention as parts of shared attention. Shared attention requires both the initiator and the responder to be aware that they are observing the same object, unlike joint attention.

Recent work showed the ability to estimate the individual's gaze target directly from images using neural networks. We differentiate between single and joint attention target estimation methods based on the number of subjects involved in this process. A key step towards single attention target estimation was the work by Recasens *et al*. [22], which demonstrated the ability to detect the attention target of each person within a single image. This image-based method did not consider human attention over time and the cases when the target of the subject's attention was outside the image frame. This approach was later extended to handle the issue of out-of-frame gaze targets [2]. Afterwards, Chong *et al*. [3] proposed a spatio-temporal approach to gaze target prediction, which models gaze dynamics from video data. These single-target estimation approaches are attractive because they can leverage head pose features and the saliency of potential gaze targets to resolve ambiguities in gaze estimation. However, unlike humans, they only use 2D information to estimate the point of interest. For the first time, Fang *et al*. [9] proposed a method using depth prior, 3D gaze estimation and 2D field-of-view (FOV) estimation for gaze target prediction. This was essential to more realistic gaze target estimation in 3D space. However, in reality, the FOV of people is not two-dimensional, as a healthy person can observe things in front of them within a 3D cone.

In social scenarios, we often infer the gaze target of two or more people simultaneously. To solve this task, an inefficient way is to use the single target estimation models and estimate the gaze target of every individual in the scene one by one and then combine these estimates to find the joint attention target of the scene. Fan *et al*. [8] proposed to infer the joint attention target in third-person social scene videos using a spatial-temporal neural network to overcome the limitations of the single target estimation methods. This solution was based on a head detector module, region proposal, and saliency estimation. Later, Sumer *et al*. [26] proposed an end-to-end solution without using any temporal information, face detector, or head pose estimator to detect and localise joint attention. Both joint attention target estimation methods rely solemnly on 2D data, making the models more prone to errors, such as physically impossible predictions where the subjects are estimated to look within their blindspot.

This study extends the previous approaches by developing a model for 3D FOV-based co-attention target estimation by jointly using 2D and 3D clues. Our motivation is to create a model that can translate the image as humans do and estimate where the subject is looking in the 3D space. Introducing strong 3D clues into this framework helps the model handle occlusion and other challenging cases better. Our contributions are fourfold:

- We propose a novel joint attention target estimation model which mimics how humans observe their environment using 2D and 3D clues.
- We trained a spatial model that can utilise the full scope of the 3D information of the 3D space provided by the monocular depth estimator. The predicted 3D FOV of the subjects are used as a probability map instead of a fixed angle, hard thresholded FOV cone to make the model more robust against the potential 3D gaze direction estimation errors.
- This is the first work to incorporate depth information into a joint attention target model and investigate its usefulness in the case of both joint and single attention target estimation tasks.
- The proposed joint attention target estimation approach outperformed the state-of-the-art single and joint attention target estimation methods. The results of the methods were compared on a large-scale image benchmark dataset and two video datasets.

In this work, we rely on an implicit social clue to infer multiple users' common gaze target point in the scene. We aim to address the physically impossible predictions of the existing models, where the subjects are predicted to observe a point within their blind spot. Inspired by the working of the human visual system, we proposed incorporating depth information into the attention target estimation pipeline. By fully utilising the depth prior generated by a monocular depth estimator module [11] combined with the subject 3D orientation, we predicted a probability map indicating the pixel-wise co-attention target likelihood on the image frame. To the best of our knowledge, this is the first work to fully utilise depth information in a joint attention target estimation framework.

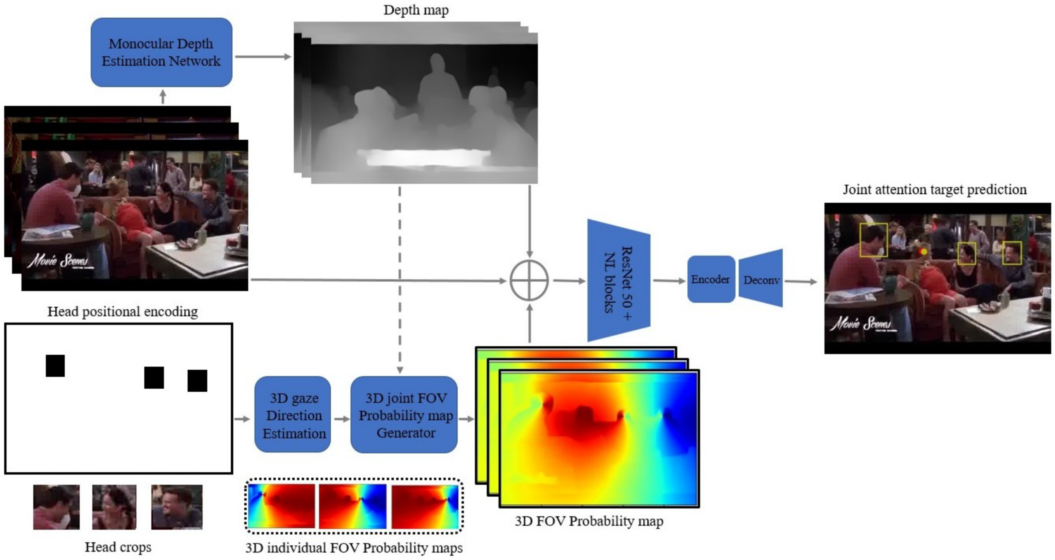
Monocular Depth Estimation Network
Depth map
Joint attention target prediction
ResNet 50 + NL blocks
Encoder
Deconv
Head positional encoding
3D gaze Direction Estimation
3D joint FOV Probability map Generator
3D FOV Probability map
Head crops
3D individual FOV Probability maps

Figure 2. **Overall framework of the proposed JAT method.** The input of the framework is the RGB image and the head bounding box annotation of the subjects of interest. Head crops of the subjects are generated based on their head positional encoding and used as the input of the 3D gaze target estimator. The depth map, generated by the monocular depth estimation network, is used as the input of the 3D field-of-view (FOV) probability map generator alongside the estimated 3D gaze direction. The generated depth and 3D FOV probability maps and the original RGB image are then inputted into the Joint attention target estimation module to predict the location of the joint attention target of the selected subjects in the scene. The ground truth attention target location is shown in yellow, and the estimate of the proposed method is in red.

2. Related Works

Attention target estimation methods can be categorised based on the number of people involved in the social interaction in third-person social images or videos.

Single attention target (SAT). This class of methods focuses on a single subject within the scene and aims to infer their visual attention target location based on the visual information. The pioneering work of this research field was proposed by Recasens *et al.* [24]. The proposed deep model was the first to learn to find the gaze target in the image through two pathways. The input of the scene saliency pathway is an RGB image designed to estimate the saliency of the scene. The subject's gaze direction was estimated through the gaze pathway, which takes the face crop of the subject and its spatial location within the original image as the input. The image dataset (GazeFollow) proposed in this paper serves as the primary large-scale benchmark of this research field. Despite the promising results presented in this work, the proposed method did not handle out-of-frame targets or modelled the temporal dynamics of attention. Chong *et al.* [3] proposed a spatio-temporal model to address these limitations. The authors proposed a video attention target dataset (VideoAttentionTarget) and extended the GazeFollow dataset with out-of-frame annotations. These methods were designed to estimate the attention target location of a single subject within the 2D image. Other related works include [13, 17, 23].

Joint attention target (JAT). Fan *et al.* [8] proposed a method to infer the joint attention target of two or more people in the scene. The method takes an image frame as input and, through a head detector, a gaze estimation module, and a region proposal module generates a joint attention spatial heatmap. Furthermore, the authors presented this task's first large-scale third-person social scene video dataset (VideoCoAtt). An end-to-end Joint attention target (JAT) estimation method was developed by Sumer *et al.* [26]. A frequent common mistake of the presented SAT and JAT methods is that they do not utilise 3D clues for scene understanding. Therefore, the target estimates are often within the subject's blind spot.

Depth-aware attention target The latest works on SAT estimation proposed by Fang *et al.* [9] and Bao *et al.* [1] addressed this limitation. [9] proposed an image-based SAT estimation model. This work does not consider temporal information, and although the authors developed a 3D gaze direction estimation module, the FOV generator only relies on 2D gaze and head direction and a view angle of 60°. [1] proposed to reconstruct the 3D scene to 3D point cloud using relative depth estimation and 3D human pose estimation. They selected the front-most 3D points along the predefined visual rays to find the final gaze target position. These works are the first and currently, only existing works which took a step towards integrating the depth clue into the attention target detection model.

The work presented in this paper is the first to use depth information for the JAT estimation task. We propose for-

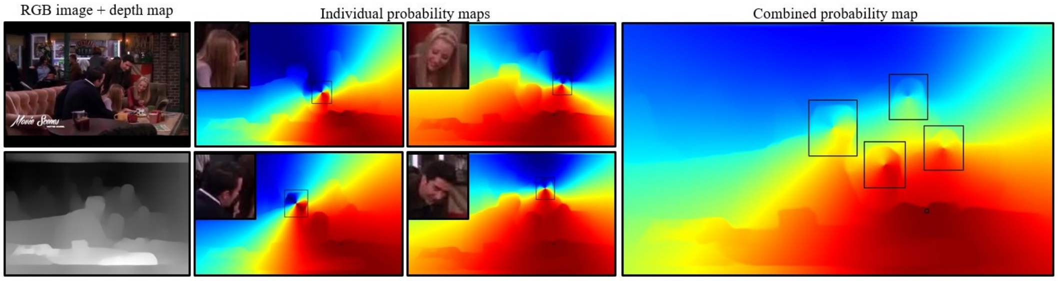

Figure 3. **Example joint 3D FOV probability map of the subjects looking at the same attention target.** We show the original input image and corresponding monocular depth map in the first column. Then we show the individual probability maps of each subject. Finally, on the combined joint attention probability map, we show the ground truth head bounding boxes and the attention target location in black.

mulating the 3D FOV as a probability map instead of hard thresholding the values along a pre-defined angle or visual rays. This new way of representing the FOV allows more room for inaccuracies in the 3D gaze direction estimator. Furthermore, our method combines the depth information with the subjects' pose to produce a joint 3D FOV probability map to predict joint gaze targets of multiple subjects within the scene.

3. Methodology

In this Section, we introduce a novel framework using a monocular depth estimator and a 3D FOV-based probability map to estimate the joint attention target while minimising the physically impossible gaze target estimate. Our assumption is that by integrating relative depth information of the scene and the calculated joint 3D FOV of the subjects in our framework, the model will learn to differentiate between the FOV of the subject and the blind spot. The framework of the proposed method is shown in Figure 2.

The input of this method is the original image, the head positional encoding. We use an existing monocular depth estimation method [11] for relative depth map generation from the single image input and a gaze direction estimation method [15] to estimate the subjects' 3D gaze direction.

3.1. Framework

Our framework comprises three major components: a *Relative Depth Prior Module*, a *3D Field-of-View Module*, and finally, a *Joint Attention Target Prediction Module*.

Relative Depth Prior Module This module is the core of the proposed method, as the generated depth map is the input of both the 3D Field-of-View and the Joint Attention Target Prediction Modules. For our task, we are primarily interested in the order of the objects and where they are located w.r.t. each other. Therefore, instead of estimating the absolute depth, we used an existing monocular depth estimation network [11] to estimate the relative depth map of the scene. Relative depth is the ratio between the depth of two points, which is useful to determine which point is closer to the camera [16].

3D Field-of-View Module The crop of the subjects' heads in the scene is used to estimate their 3D head orientation using an existing 3D gaze estimator module. The 3D direction estimate combined with the 2D spatial positional encoding of the head bounding boxes allows us to generate the subjects' 3D individual FOV (shown at the bottom of Figure 2). We generate a shared 3D FOV probability map for each image, including every subject. Based on the assumption that a person is more likely to look within their 3D FOV cone than to their blind spot, we assigned a higher probability for the joint attention targets to be within the intersections of the subject's 3D FOV cones, and we penalise the predictions which fall outside of the cones. In addition, based on our preliminary experimental results presented in Section 1. of the Supplementary material, we assigned the lowest probability score to the subjects' head bounding box region. The generator outputs are joint 3D FOV probability maps corresponding to the input images. The individual probability map generation is mathematically denoted as:

$$M_{ind} = min\_max\_scaler\left(\frac{(i-h_x,j-h_y,k-h_z)\cdot(g_x,g_y,g_z)}{\|i-h_x,j-h_y,k-h_z\|_2\cdot\|g_x,g_y,g_z\|_2}\right), \quad (1)$$

where ind={0,...,n} is the index of the subjects in the scene, (i,j,k) is the coordinate of each point in M_{ind},(h_x, h_y, h_z) is the centre of the head bounding box, (g_x, g_y, g_z) is the estimated 3D gaze direction, and min_max_scaler() is the transformation that scales each value of the probability map between zero and one. Then we set the values within the subject's head bounding box ($[x_{min}, x_{max}, y_{min}, y_{max}]$) to be equal to zero.

$$M_{ind}[x_{min}, x_{max}, y_{min}, y_{max}] = 0 \quad (2)$$

Finally, the joint attention probability map values are calculated as the average of the individual probability maps.

$$M_{FOV} = mean(M_{ind}) \quad (3)$$

An example visualisation of the generated individual and joint 3D FOV probability maps is shown in Figure 3. Four subjects are in the image, their head crops are highlighted at the left top corner of the individual heatmaps, and the positional encoding is visualised as black bounding boxes on the joint probability map. The ground truth JAT point is visualised in black on the joint probability map. We can see that the highest probability map values correspond to the area where the ground truth point is located within the scene. Furthermore, by using the relative depth prior information (shown at the bottom left of the figure), we can see the clear difference between the pixel values of the blind spot of the individuals and the FOV.

Joint Attention Target Prediction Module Finally, we defined a JAT prediction module to localise the attention target point of the individuals in the scene. The input of this module is a series of scene images, the corresponding relative depth maps, and the calculated 3D joint FOV probability maps. These inputs are concatenated and fed into an encoder consisting of a ResNet-50 [14] followed by an additional residual and average pooling layer combined with NL layers [29] In between the layers of the second and third residual blocks, we included 3 and 5 NL layers, respectively. The concatenated features are encoded using two convolutional layers in the Encoder. A deconvolutional network composed of four deconvolution layers upsamples the features calculated by the Encoder into a full-sized feature map. We found that by combining the scene and the subject-dependent information using these inputs, we can find the most probable gaze target location on the image from the joint FOV of the subjects.

3.2. Implementation details

We implement our method in PyTorch. All experiments are run on an Intel i9-CPU @ 3.30GHz, 125 GB RAM, and four NVIDIA GeForce RTX 2080Ti GPUs.

The input RGB and generated depth images are resized to 224×224 and normalised. As described in [24], we used random flip, colour jitter, and crop augmentations. We also added noise to the head position and the 3D gaze direction during training to minimise the influence of localisation errors. The ground truth heatmap was generated using the output 3D direction estimate calculated by [15] and by adding Gaussian weight around the centre of the target for supervision. We implemented two loss functions during training: heatmap and in-frame loss. We used MSE loss to compute the heatmap loss ($\mathcal{L}_h$) and binary cross-entropy loss for the in-frame loss ($\mathcal{L}_f$). The total loss $\mathcal{L}$ used for training is a weighted sum of these two: $\mathcal{L} = w_h \mathcal{L}_h + w_f \mathcal{L}_f$.

4. Experiments

4.1. Dataset and baselines

We performed experiments to evaluate the proposed JAT estimation method on the GazeFollow [24] image, and VideoAttentionTarget [3] video SAT benchmark datasets. The proposed JAT method was compared to the following state-of-the-art SAT models' performance: solutions utilising depth prior **DAM [6]**, **ESCNet [1]** and others without **VideoAttentionTarget [3]**, and **HGTTR [28]**.

Furthermore, we evaluated the performance of our method on the social interaction detection VideoCoAtt [8] JAT video benchmark dataset. On the JAT estimation task, we compare our method against the following methods: **Fan [8]** the first method proposed to infer joint attention in social scene videos, **Sumer [26]** an end-to-end method designed for JAT estimation on videos, **VideoAttentionTarget [3]** a single attention target estimation method used on videos, and **Attention Flow [26]** an End-to-End Joint Attention Estimation method. Our method produces state-of-the-art results on all datasets in all experiments.

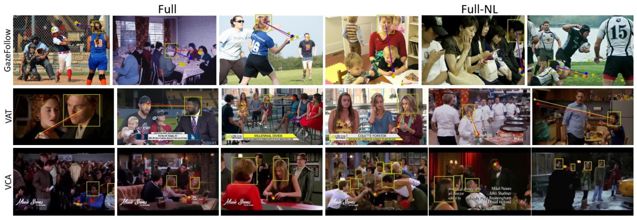

Figure 4. **Qualitative highlights of the proposed method with NL layers (Full-NL) and without (Full)** on three attention target estimation datasets: GazeFollow, VAT, and VCA. In the presented examples, the head bounding box of the observed subjects and the ground truth annotations are marked as yellow, the average is shown in blue, and the estimated gaze target estimate is in red.

4.2. Comparison with the state-of-the-art

For the most exhaustive comparison, the proposed joint attention target estimation model is evaluated and compared against both single and joint attention target estimation methods. We present our experiments' quantitative and qualitative results using three aforementioned benchmark datasets.

4.2.1 Qualitative results

The qualitative highlights of the proposed Full and Full-NL methods on three benchmark datasets, GazeFollow, VAT and VCA, are shown in Figure 4. These examples were selected to demonstrate the efficiency of our method in different scenarios, *e.g.* in case of occlusion and ground truth ambiguity. Furthermore, we selected cases when the gaze target was not another person in the scene to address the human bias problem. See the detailed discussion on the failure cases of the previous attention target estimation methods in Section 2. of the Supplementary material.

The attention target estimate of our methods is shown in red, the ground truth annotations and the head bounding

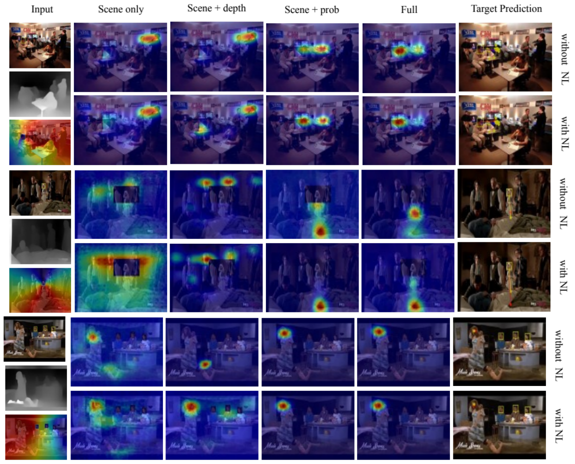

Figure 5. **Qualitative results of ablation study on the GF SAT image, VAT SAT video and the VCA JAT video benchmark datasets, respectively.** The input of the Full and Full-NL proposed models and their variants, the RGB image, the generated prior depth map and the corresponding calculated 3D FOV probability map are shown in the first column. We visualised the generated output heatmap of every variant (Scene only, Scene+depth, Scene+prob) and the Full method (Scene+depth+prob) and, finally, the gaze target prediction of the Full method. In the target prediction visualisation and the input image, the head bounding box of the observed subjects and the ground truth annotations are marked as yellow, and the estimated gaze target is shown in red.

boxes of the observed subjects are yellow. For the GazeFollow dataset, we show the average ground truth annotation location in blue. These examples show that even in challenging cases *e.g.* when the subject's face is occluded or not visible or when the person is surrounded by many people around them and is looking at an object or in the middle of the field, our models successfully identified the attention target location.

Additional qualitative results are shown in the last two columns of Figure 5. In the first column of this figure, we show the original input image, the prior depth map and the 3D FOV probability map for each example. Following these in the penultimate column, we show the predicted heatmap of Full and Full-NL. In the last column, we visualised the output target prediction, the ground truth annotation and the subject's head bounding box. We observed that the two models produce similar heatmaps reflected in the calculated AUC scores presented in Section 4.2.2. We found that the Full-NL model generated more confident and correct heatmaps when the scenario was complex (See the first examples from the GazeFollow dataset). This might be due to the NL layers' ability to capture long-range spatial dependencies. For less complicated cases, *e.g.* the second example from the VAT dataset shown in Figure 5, we observed that the use of the NL layers introduced unwanted dependencies. Additional qualitative results are shown as part of our Supplementary material in Section 4.

Table 1. **Quantitative evaluation and ablation study results and comparison with the state-of-the-art methods on the GazeFollow SAT estimation image dataset.** Gaze direction estimation error shows the range of the random noise added to the 3D gaze direction of the subjects before the probability map generation.

Method	AUC ↑	Min dist. ↓	Avg dist. ↓
Scene only	0.889	0.143	0.213
Scene + depth	0.894	0.136	0.205
Scene + prob	0.928	0.036	0.084
Full (scene + depth + prob)	**0.932**	**0.036**	**0.082**
NL + Scene only	0.883	0.148	0.216
NL + Scene + depth	0.894	0.136	0.204
NL + Scene + prob	0.925	0.033	0.082
Full-NL (NL + scene + d + prob)	**0.926**	**0.028**	**0.075**
Full gaze dir error ± 13.5°	0.930	0.052	0.100
Full gaze dir error ±30°	0.927	0.047	0.097
Full-NL gaze dir error ±13.5°	0.932	0.039	0.087
Full-NL gaze dir error ±30°	0.929	0.049	0.099
HGTTR [28]	0.905	0.065	0.138
VideoAttention [3]	0.921	0.077	0.137
DAM [9]	0.922	0.067	0.124
ESCNet [1]	0.928	-	0.122

4.2.2 Quantitative results

Here, we present the results of the quantitative evaluation. Note that the evaluation metrics differ for each benchmark dataset. For more details on the datasets and metrics, refer to Section 3 of the Supplementary material. The type of attention target estimation tasks and benchmark datasets organise the results in this section.

SAT estimation on the GazeFollow dataset. The quantitative results on the GazeFollow dataset are shown in Table 1. We compared the performance of our method in the SAT estimation task with the latest methods HGTTR [28], VideoAttention [3], DAM [9], and ESCNet [1] on this image benchmark dataset. Among these, HGTTR, DAM and ESCNet were specifically designed to solve this task and similar to our solution, DAM and ESCNet used partial, relative depth prior information in their method. The results highlighted in Table 1 show that in terms of all the evaluation metrics, the proposed Full-NL framework outperformed all the existing methods. Even though our framework is not designed for SAT, 58.20% minimum distance compared to [9] and 37.33% average distance compared to [1] relative improvement was achieved by our method on the GazeFollow dataset.

Note that the performance of the proposed method with (Full-NL) or without (Full) the additional NL layers is very similar. We found that in the presence of relative depth prior information, the NL layers did not contribute towards the performance significantly.

SAT estimation on the VAT dataset. Furthermore, we compared our solution with the same methods on the SAT

Table 2. **Quantitative evaluation and ablation study results and comparison with the state-of-the-art methods on the VAT SAT estimation video dataset.** Gaze direction estimation error shows the range of the random noise added to the 3D gaze direction of the subjects before the probability map generation.

Method	AUC ↑	L2 dist. ↓
Scene only	0.711	0.306
Scene + depth	0.728	0.313
Scene + prob	0.935	0.082
Full (scene + depth + prob)	**0.937**	**0.077**
NL + Scene only	0.713	0.318
NL + Scene + depth	0.743	0.334
NL + Scene + prob	0.944	0.082
Full-NL (NL + scene + depth + prob)	**0.951**	**0.074**
Full gaze dir error ± 13.5°	0.930	0.134
Full gaze dir error ± 30°	0.911	0.122
Full-NL gaze dir error ± 13.5°	0.943	0.093
Full-NL gaze dir error ± 30°	0.914	0.153
ESCNet [1]	0.885	0.120
VideoAttention [3]	0.860	0.134
HGTTR [28]	0.904	0.126
DAM [9]	0.905	0.108

Table 3. **Quantitative evaluation and ablation study results and comparison with the state-of-the-art methods on the VCA JAT estimation video dataset.**. Gaze direction estimation error shows the range of the random noise added to the 3D gaze direction of the subjects before the probability map generation.

Method	L2 dist. ↓	Pred. Acc. ↑
Scene only	139	13.0
Scene + depth	130	41.5
Scene + prob	16	90.0
Full (scene + depth + prob)	**13**	90.2
NL + Scene only	145	17.1
NL + Scene + depth	145	31.8
NL + Scene + prob	14	93.0
Full-NL (NL + scene + depth + prob)	**13**	**93.2**
Full gaze dir error ± 13.5°	21	83.5
Full gaze dir error ± 30°	24	50.1
Full-NL gaze dir error ± 13.5°	19	80.8
Full-NL gaze dir error ± 30°	34	66.2
Fan [8]	62	71.4
Sumer [26]	63	78.1
VideoAttention [3]	57	83.3
HGTTR [28]	46	90.4

estimation using the VAT video benchmark dataset. The results are shown in Table 2. We compared our performance with the previously mentioned HGTTR [28], VideoAttention [3], DAM [9] and ESCNet [1] methods. Both proposed methods were more efficient at estimating the gaze target than the state-of-the-art methods. We found that Full-NL outperformed the performance of the Full method in terms of both AUC and L2 distance measures. Full-NL improved the AUC score by 4.84 % and the L2 distance by 31.48 %.

JAT estimation on the VCA dataset. Finally, the quantitative results on the VCA video benchmark dataset are shown in Table 3. We compared the JAT estimation performance of our method with Fan [8], Sumer [26], VideoAttention [3], and HGTTR [28]. Among these state-of-the-art methods, Fan, Sumer and HGTTR were trained to estimate the attention target location of multiple subjects in the scene. The results showed that the proposed method with and without the NL layers significantly outperformed all the state-of-the-art methods in terms of the L2 distance metric. We were able to reduce the distance error by 71.74 % compared to the result report by [28]. The NL layers proved useful in achieving the best prediction accuracy. Overall, the proposed method achieved state-of-the-art performance in terms of all evaluation metrics on this dataset too.

In summary, the quantitative results confirmed that the JAT estimation method proposed in Section 3 achieved state-of-the-art performance across all the benchmark datasets and their evaluation metrics on the SAT and JAT estimation tasks. Furthermore, the comparison between Full and Full-NL across the datasets shows that the usefulness of the NL layers is context and complexity dependent.

4.3. Ablation Study

To study the contribution and effectiveness of different components of the proposed method, we trained several models with different parameters. In this section, we discuss the findings of these experiments on three benchmark datasets.

4.3.1 Spatial model components

We trained the following variations of the proposed full spatial method: Scene only, Scene+depth, Scene+probability map, Scene+depth+probability map, and their variants, including the non-local layers in the encoder. Qualitative highlights are shown in Figure 5. Note that the observations discussed below are accurate for all the benchmark datasets.

Across all the benchmark datasets, we found that the Scene only variant performed the worst compared to the other variants. The heatmaps in the second column of the qualitative highlights figures also confirmed that the predicted output heatmap of this module alone, most of the time, did not overlap with the gaze target area of the image, and it was widespread and not confident. This is because the model was unaware of any subject-specific information; therefore, it relied solemnly on the scene information to estimate the subject-dependent attention target location.

We also found that when we combined the scene information with the output of the prior depth map of the monocular depth estimator, the performance of the trained Scene+depth models improved slightly. The estimated output heatmaps of these models (See the third column of qualitative figures) were more successful in localising the FOV of the subject. These heatmaps were more confident; however, they often misplaced the gaze target as it was selected based on the Scene information. Therefore, despite this improvement, as the input of these models was still subject-independent, their results were not satisfactory.

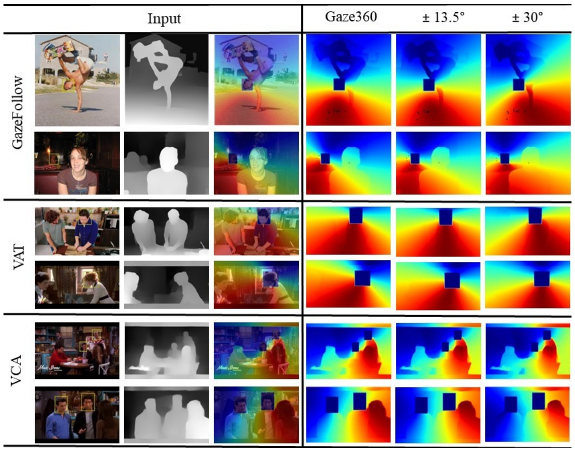

Figure 6. **Visualisation of the joint 3D FOV probability map effected by gaze direction error.** We show the generated probability map using the 3D gaze estimator Gaze360 [15] and when additional 13.5 and 30 degree gaze direction error was added.

The proposed 3D FOV probability map contains depth and subject-dependent information. Introducing subject-dependent information into the model significantly improved the performance quantitatively and qualitatively. The output heatmaps of Scene+prob, shown in the fourth column of Figure 5, are concentrated around the correct gaze target location, the location of the maximum of these heatmaps shifted significantly from the predictions of the Scene only and Scene+depth models.

Finally, we can see that explicitly using the prior depth map as an input and not just as a part of the probability map further improved the results. While the improvement was moderate compared to the Scene+prob performance. However, the Scene+depth+prob with (Full-NL) or without (Full) the NL layers proved to be the most efficient in estimating the attention target of single or multiple subjects within the scene.

In summary, the ablation study confirmed that all the modules included in the Full model (Scene+depth+probability map) are useful and contribute to the proposed solution's performance. We demonstrated that relying only or too much on the scene information is insufficient to estimate the subject's gaze target location.

4.3.2 Gaze direction estimation error

The proposed probability map's role in the outstanding performance of the proposed method has been demonstrated through the previous experiments. The input of the 3D FOV probability map is the 3D gaze target estimate of the observed subject. To test the robustness of the proposed method against gaze direction prediction errors, we trained two variants of the Full and Full-NL models under extreme error levels.

During this experiment, we generated the 3D FOV probability map using additional random noise added to the subjects' estimated gaze direction. We chose the noise levels to reflect the average error (± 13.5°) of the state-of-the-art 3D gaze direction estimation method [15] and to reflect the human's horizontal central vision range (± 30°). We show example probability map variants generated with additional gaze direction error in Figure 6.

While adding NL layers to the proposed method did not improve the performance significantly under moderate gaze estimation error in the previously presented experiments, our results show that the Full-NL models were more robust against the additional noise than the Full models. Furthermore, we found that in the case of the large-scale GazeFollow image dataset (See Table 1) and the VAT video dataset (See Table 2) the proposed model surpassed the performance of the state-of-the-art methods even when we added ± 30° gaze direction estimation error, which is more than double the existing 3D gaze estimators' average angular error. These results on the SAT estimation task are especially outstanding as the proposed 3D FOV probability map is the most useful in improving the robustness of the attention target estimation when there is more than one subject within the scene.

5. Conclusion

In this paper, we proposed a novel joint attention target estimation framework which was developed to fully utilise the 3D clues of the scene efficiently. Following the findings of our preliminary experiments, we aimed to tackle the human bias and physically impossible predictions, which are the major flaws of the previously proposed models. To achieve this, we proposed to combine a novel 3D field-of-view-based joint attention probability map with the scene and depth information. Extensive qualitative and quantitative analysis on three benchmark datasets shows that the proposed method achieved favourable performance compared to both the state-of-the-art single and joint attention target estimation approaches. The demonstrated outstanding performance of the proposed method proved our hypothesis that using 3D clues for the third-person view attention target estimation is advantageous.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2022-0-00608, Artificial intelligence research about multi-modal interactions for empathetic conversations with humans (50%); No.2021-0-00034, Clustering technologies of fragmented data for time-based data analysis (50%)).

References

[1] Jun Bao, Buyu Liu, and Jun 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. 3, 5, 6, 7

[2] 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. 2

[3] 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, 5, 6, 7

[4] Meir Cohen, Ilan Shimshoni, Ehud Rivlin, and Amit Adam. Detecting mutual awareness events. *IEEE transactions on pattern analysis and machine intelligence*, 34(12):2327–2340, 2012. 1

[5] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. 1

[6] Huiyu Duan, Xiongkuo Min, Yi Fang, Lei Fan, Xiaokang Yang, and Guangtao 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. 1, 5

[7] Nathan J Emery. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & biobehavioral reviews*, 24(6):581–604, 2000. 1, 2

[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. 2, 3, 5, 7

[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*, pages 11390–11399, 2021. 2, 3, 6, 7

[10] Chris D Frith and Uta Frith. Social cognition in humans. *Current biology*, 17(16):R724–R732, 2007. 2

[11] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019. 2, 4

[12] Siavash Gorji and James J Clark. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2510–2519, 2017. 2

[13] 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. 3

[14] 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

[15] 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. 4, 5, 8

[16] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019. 4

[17] 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. 3

[18] 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*, pages 3477–3485, 2019. 1

[19] Manuel Jesús Marin-Jimenez, Andrew Zisserman, Marcin Eichner, and Vittorio Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014. 1

[20] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 1

[21] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998. 1

[22] 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. 2

[23] 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. 3

[24] Adria Recasens Continente, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Neural Information Processing Systems Foundation*, 2015. 3, 5

[25] Lisa J Stephenson, S Gareth Edwards, and Andrew P Bayliss. From gaze perception to social cognition: The shared-attention system. *Perspectives on Psychological Science*, 16(3):553–576, 2021. 2

[26] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3327–3336, 2020. 2, 3, 5, 7

[27] Michael Tomasello, Brian Hare, Hagen Lehmann, and Josep Call. Reliance on head versus eyes in the gaze following of

great apes and human infants: the cooperative eye hypothesis. *Journal of human evolution*, 52(3):314–320, 2007. 2

[28] Danyang Tu, Xiongkuo Min, Huiyu Duan, Guodong Guo, Guangtao Zhai, and Wei Shen. End-to-end human-gaze-target detection with transformers. *arXiv preprint arXiv:2203.10433*, 2022. 5, 6, 7

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 5