Multimodal Across Domains Gaze Target Detection

Francesco Tonini
Department of Information Engineering and Computer Science, University of Trento
Trento, Italy
francesco@tonini.dev

Cigdem Beyan
Department of Information Engineering and Computer Science, University of Trento
Trento, Italy
cigdem.beyan@unitn.it

Elisa Ricci
Department of Information Engineering and Computer Science, University of Trento
Trento, Italy
Deep Visual Learning Research Group, Fondazione Bruno Kessler
Trento, Italy
e.ricci@unitn.it

ABSTRACT

This paper addresses the gaze target detection problem in single images captured from the third-person perspective. We present a multimodal deep architecture to infer where a person in a scene is looking. This spatial model is trained on the head images of the person-of-interest, scene and depth maps representing rich context information. Our model, unlike several prior art, do not require supervision of the gaze angles, do not rely on head orientation information and/or location of the eyes of person-of-interest. Extensive experiments demonstrate the stronger performance of our method on multiple benchmark datasets. We also investigated several variations of our method by altering joint-learning of multimodal data. Some variations outperform a few prior art as well. First time in this paper, we inspect domain adaptation for gaze target detection, and we empower our multimodal network to effectively handle the domain gap across datasets. The code of the proposed method is available at https://github.com/francescotonini/multimodal-across-domains-gaze-target-detection.

CCS CONCEPTS

• **Applied computing**; • **Computing methodologies** → **Machine learning approaches**; • **Human-centered computing**;

KEYWORDS

Gaze target detection; gaze following; domain adaptation; RGB image; depth map; multimodal data

ACM Reference Format:
Francesco Tonini, Cigdem Beyan, and Elisa Ricci. 2022. Multimodal Across Domains Gaze Target Detection. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3536221.3556624

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556624

1 INTRODUCTION

Gaze behavior indicates the visual attention of a person and allows to specify what a person is interested in, helps to decipher and forecast the interactions, intentions or actions of people [10, 32, 40]. Human-beings have a remarkable capability to detect the gaze direction of others, understand whether a person is gazing them, follow other's gaze to identify their target, and determine the attention of others [4]. However, automatically performing and quantifying these remains as a challenging problem. The research on automatic gaze behavior analysis is divided as *gaze estimation* and *gaze target detection* [5, 10, 17, 36]. Gaze estimation refers to determining the person's gaze direction (typically in 3D) while does not focus on accurately locating where a person in the scene is looking [25, 49, 57]. Instead, gaze target detection (also referred as *gaze-following* [5, 10, 18]) is to inferring where each person in the scene (2D or 3D) is looking [25, 36, 40]. This paper addresses the ***gaze target detection*** in ***single images*** (i.e., in 2D), collected ***in-the-wild***, and captured from the ***third-person perspective***. In this scope, earlier works present Convolutional Neural Network (CNN)-based architectures composed of two-pathways. While one path learns feature embeddings from the scene images, the other one models the head image belonging to the person whose gaze target is to be predicted [23, 35, 36]. Studies [23, 35, 36] perform spatial modeling, instead Chong et al. [5] extended the aforementioned two-pathway architecture by explicitly modeling the embeddings of the scene and head images over time (i.e., apply spatio-temporal modeling). That method [5] presents improved results with respect to earlier research, however, it still lacks of understanding the so-called person-relative depth. As a consequence, false detections occur when there are multiple object-of-interests at different depths but along with the subject's gaze direction. This handicap was handled in [10, 18] by integrating the depth images into the pipeline. Fang et al. [10] additionally relies on head pose detection, eyes detection and eye features extraction. Such a framework [10] improved the gaze target detection performance, while potentially being error-prone in real-life processing, e.g., when the eyes are not visible or detectable. On the other hand, Jin et al. [18] involves an auxiliary network to perform 3D-gaze orientation estimation using pseudo labels, in addition to using another auxiliary network to estimate the depth. The performance of [18] depends on reliable depth and orientation pseudo labels.

Unlike [23, 35, 36], we do not require supervision of gaze angles, which simplifies our training process and improves its applicability.

Different from [5], we apply only spatial processing, but still able to detect the gaze target at each frame of a video. Similar to [10, 18], we use depth images. Our ***multimodal pipeline*** (see Sec. 3.1) has three-pathways to process: *i)* the head image, *ii)* the scene image and *iii)* the depth map, which is obtained by standalone monocular depth estimation from RGB images [34]. It is important to highlight that our pipeline is computationally low-cost and simpler than [10] by not requiring the detection of the head pose and the location of the eyes. Furthermore, unlike [18], our proposal does not use an additional network to estimate gaze orientation from head features. Instead, we implicitly learn the orientation features using the head attention module. Given the proposed pipeline, one major aspect of this paper is to investigate how different modalities should be jointly learned for performing effective gaze target detection. To do so, we present a comprehensive experimental analysis (see Sec. 4.2). Consequently, not only the proposed method but also some of the variations of it exceed the performance of the prior art.

Generalization capability of a trained gaze target detection model is of paramount importance for its utilization in practice. However, empirical analysis (Sec. 4.4) show that, the performance of a gaze target detection model significantly decreases when it is tested on a dataset different from the one it is trained on. This phenomena is equally valid when the training is performed on the in-the-wild datasets (Sec. 4.1). Motivated by this, as the first attempt in the gaze target detection literature, this paper studies the domain-shift problem, and propose a novel ***domain adaptation method*** integrated into the proposed multimodal gaze target detection architecture (Sec. 3.2). Our method improves the results remarkably by also outperforming a state-of-the-art (SOTA) domain adaptation method. The main contributions of this study can be summarized as follows.

- A novel multimodal deep architecture, that detects the gaze target in a 2D-image captured from the third-person perspective, is proposed.
- We empirically validate the performance of the proposed model on several benchmark datasets in which it shows improved results relative to the SOTA.
- We show the effectiveness of the proposed model through an ablation study and by presenting several variations of it. Even some of the variations achieve better scores compared to the SOTA.
- This work is the first attempt where the domain adaption for gaze target detection is studied. We first diagnosed the domain-shift problem for our model as well as the prior art, and then propose a novel method to handle it.
- The proposed domain adaptation approach results in enhanced performance on target datasets, which is also superior than a SOTA multimodal domain adaptation method.

2 RELATED WORK

Below, we describe the related studies for gaze target detection task. Then, we briefly summarize the domain adaptation (DA) research in general, and focus on DA for gaze behavior analysis and multimodal visual data.

2.1 Gaze Target Detection

Gaze target detection has applications in several fields, e.g., human interaction systems, computer vision and robotics, where it is important to understand the object-of-interest [37], predict and anticipate the actions [22, 30] and so forth. Most existing works on gaze target detection rely on a particular sensor (eye trackers [40], VR/AR devices [8], RGB-D cameras [17, 47], etc.) or applicable for specialized settings (e.g., face-to-face meetings [1]) or applications (e.g., identifying the mutual gaze [27], detecting the common gaze point of multiple human observers [53, 60], anticipating averted gaze [31]) or requires constrained subject placement in the scene [29]. Another categorization is regarding whether the target is in 2D image [4, 5, 23, 35, 36] or 3D space [3, 17, 28, 47].

In this paper, we focus on the gaze target detection in *2D-single images* which are collected in *unconstrained environments* from the *third-person view*. In this context, one of the first work adapting deep learning architectures was [35], which present two-pathway architecture. One branch of that network [35] takes the scene images to estimate the saliency (so-called saliency pathway) while the other one (so-called gaze pathway) gets head images as the input and models the gaze direction. An effective component of that network [35] is the head location information injected into the gaze pathway, which improves the gaze target detection results remarkably. The posterior work [4] adapted the aforementioned two-pathway architecture while others [5, 10, 18, 23, 36] utilized both the two-pathway model and the head information injection pipeline. Differently, Chong et al. [4] extended the architecture of [35] to detect the gaze targets not-being in the scene (so called out-of-frame gaze targets) by simultaneously learning the gaze angles and the saliency. The out-of-frame component (*two convolutional layer + ReLU + softmax*) is kept the same in [5, 10]. Chong et al. [5] additionally integrated CNN-LSTM, which processes the feature embeddings of the gaze and saliency pathways to learn the gaze behavior in time. Even though [5, 23, 35, 36] present promising results for gaze target detection, they all fail to address the challenge of handling the situations where the person-relative depth matters. For instance, in the situations where there are multiple objects at different depths along with the subject's gaze direction, it is unlikely that the correct gaze target can be determined by any of these methods. In order to handle this, [10] utilize the depth information and 3D-gaze to produce target-focused spatial attention map. However, the overall pipeline of [10] is computationally heavy as it requires detection of head pose and eyes. Jin et al. [18] is another study which include the depth maps. The authors [18] designed a primary network that predicts gaze as in [5]. Furthermore, to improve the prediction performance, they introduce two auxiliary networks: one to learn depth features, the other to learn 3D-gaze orientation features. The whole pipeline is learnt using the ground-truth, pseudo depth and orientation labels.

Our work diverges from prior art in several aspects. First of all, unlike [23, 35, 36], we do not require supervision of gaze angles. Different from [5, 23, 35, 36], we adapt the depth images (obtained by monocular depth estimation method [34]) and consequently improve the spatial modeling by handling the person-relative depth challenge. Unlike [5], we not only apply spatial pooling in the scene network by using the head features that supplies a regulation through attention mechanism, but also integrate this for the depth network, resulting in improved performance. Also, we only rely on the spatial information, and do not apply spatio-temporal data processing (i.e., less training time with respect to [5]). We present a

three-pathway network and joint learning with late fusion of head location regulated, scene and depth feature embeddings, without using head poses and location of eyes as applied in [10]. Unlike [18] our work neither requires additional depth and orientation pseudo labels nor additional networks to explicitly learn a 3D orientation representation.

We adapted the out-of-frame component of [5], but this is performed attached to the multimodal framework, which is not the case in [5, 10]. Our framework achieves better performance compared to prior art, and in some cases even surpasses the human performance. Importantly, this is the first work investigating the domain-shift problem for gaze target detection in 2D images, and presenting a relatively simple but effective multimodal DA method to boost the generalizability of the proposed three-pathway network.

2.2 Domain Adaptation

Unsupervised Domain Adaptation (UDA) is a broadly investigated methodology in order to handle the problems coming out due to the domain gap, which can happen when the training and testing data are belonging to different distributions. UDA transfers knowledge from a labeled dataset (called source domain) to another domain (called target domain), whose data is available at training time but *without labels* [48]. The literature of UDA can be divided into three as: *i)* discrepancy-based techniques [26, 52, 56] that try to minimize the distance between source and target distributions at feature level, *ii)* adversarial methods [6, 59] having a generator and a discriminator and trying to have features created by the generator as close as possible to those of the source, and *iii)* self-supervised methods [7, 45, 51] optimizing the (self-supervised) objective function to produce robust representations for the main task.

There exist relatively few research addressing to adapt to unseen domains for gaze estimation task, while to the best of our knowledge, this has not yet been investigated for the gaze target detection in 2D images (i.e., the aim of this paper). For the gaze estimation problem, Kellnhofe et al. [20] adapts the adversarial discriminative DA of [43] in which a discriminator identifies the source domain of the image features as a binary classification task in addition to having another loss used to exploit the left-right symmetry of the gaze-estimation task, providing the consistency on unlabeled data by computing the gaze of the original and horizontally flipped images in order to minimize the angular difference among two. Yu et al. [55] approach the gaze adaptation problem in terms of gaze redirection given that the eye structures of different persons cause a domain gap resulting in poor performance. To handle this, authors [55] generate synthetic eye images from existing reference samples (i.e., self-supervised DA) and define the gaze redirection loss (calculated based on enforcing that the gaze predicted from gaze redirected image is close to its target ground-truth) in addition to cycle consistency loss of [59]. Recently, Guo et al. [14] present UDA gaze estimation by embedding with prediction consistency, which ensures that linear relationships between gaze directions in different domains remain consistent on gaze space and embedding space.

On the other hand, for multimodal visual data DA, most of the prior art considers single modality. A major number of work investigated the domain-shift across RGB images while depth images were left behind. There exist methods [15, 16, 38] defining RGB images as source, and depth images (obtained from RGB-D camera) as target. Xiao et al. [21] defines the RGB-D as source while RGB as the target. Instead [2, 44] use the depth information as the additional channel for source and target. Unlike aforementioned litreature, [11] recently presents a self-supervised modality translation, showing the SOTA results for RGB-D scene recognition. Ferreri et al. [11] defines two encoder-decoder architecture based on ResNet-18; one for RGB branch and the other for depth branch. The depth decoder reconstructs the RGB image when the encoder's output is the depth embeddings, and the RGB decoder reconstructs the depth image when the encoder's output is the RGB embeddings. Authors also use an additional ResNet-18 to regulate the content similarity between the reconstructed images and the original images. In this paper, we adapt the method of [11] into our multimodal network to perform gaze target detection, whose results are compared with the proposed DA method.

The proposed DA method conducts in three networks: head, scene and depth, simultaneously. We inject a Gradient Reversal Layer (GRL) [13] between our head backbone and a domain classifier we define (which decides whether a head image belongs to the source or the target domain). We also apply RGB→Depth and Depth→RGB modality translations by attaching additional decoders to our scene and depth backbones.

3 METHODOLOGY

Given an RGB image $\mathbf{S}_i$ of the scene, the depth map $\mathbf{D}_i$ obtained from $\mathbf{S}_i$ using the state-of-the-art (SOTA) monocular depth estimator [34], $\mathbf{H}_i$; a region of interest in $\mathbf{S}_i$ which contains the head of the person (i.e., the gaze source), and a binary head location mask $\mathbf{M}_i$ in the size of $\mathbf{S}_i$ when $M_i(x, y) = 1$ for each pixel of the person's head, the aim of our architecture is to generate a 2D heatmap $\mathbf{Heat}_i$ in the size of $\mathbf{S}_i$ when the higher values are closer to the ground-truth gaze coordinate and $\underset{x,y}{\operatorname{argmax}}\ Heat_i(x, y)$ is the gaze target in pixel coordinates. In addition to $\mathbf{Heat}_i$, our neural network also has the output $\mathbf{InOut}_i$ such that $\mathbf{InOut}_i = 1$ if and only if the gaze target is inside the frame.

Our proposal builds on the work of Chong et al. [5] and injects the depth modality, which together with RGB modality supplies a richer representation of the scene, and consequently performs better in the challenging scenarios. The proposed network is composed of multiple self-contained modules. The ***scene and depth networks*** (**SN** and **DN**) process $\mathbf{S}_i$ and $\mathbf{D}_i$, respectively. The ***head network*** (**HN**) processes $\mathbf{H}_i$ independently, and produces an attention map that is then multiplied by the embeddings of **SN** and **HN**. The ***fusion and prediction module*** (**F&P**) concatenates the final scene, depth, and head features to obtain the outputs of our proposal: $\mathbf{Heat}_i$ that encodes the region in which gaze happens, and $\mathbf{InOut}_i$; the probability of the gaze target being inside or outside the scene. Different from [5], we also introduce the channel-wise outer product [39] and summation operators for multimodal joint learning. Furthermore, we bring in a ***domain adaptation* (DA) module** that relies on a Gradient Reversal Layer [13] on **HN** as

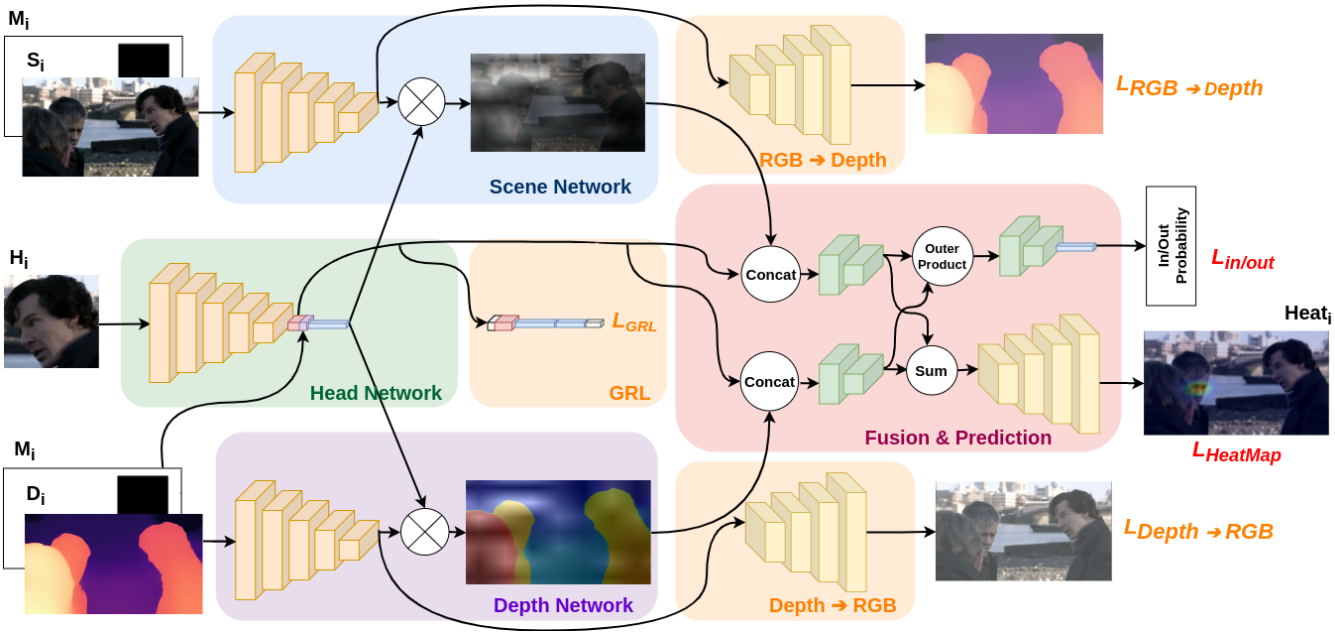

Figure 1: Our three-pathway architecture composed of scene network (blue box), head network (green box) and depth network (purple box). The input of these networks, respectively, are: *i)* scene image (S_i) concatenated with head location mask (M_i; a binary image, the bounding box of the person whose gaze to be predicted is black, the rest is white), *ii)* head image (H_i) belonging to the person whose gaze to be predicted, and *iii)* depth map (D_i) obtained by [34] concatenated with M_i. We also present the intermediate images obtained by scene and depth networks, where the effect of head attention can be seen. The multimodal joint learning and gaze prediction are performed by the fusion and prediction component (pink box). The outputs of our architecture are: *iv)* 2D heatmap ($Heat_i$, superimposed on the scene image, associated loos is $L_{HeatMap}$) and *v)* the probability of the gaze being inside or outside the scene (loss shown as $L_{in/out}$). We also perform domain adaptation (yellow boxes) *vi)* attached to head network (shown as GRL, with the loss: L_{GRL}), *vii)* attached to depth network to perform adaptation from depth to RGB images (shown as Depth→RGB, with loss: $L_{Depth \to RGB}$), and *viii)* attached to scene network to perform adaptation from RGB to depth (shown as RGB→Depth, with the loss: $L_{RGB \to Depth}$).

well as two modality decoders that work on **SN** and **DN**, respectively. An illustration of our network is given in Fig. 1. The code is available at https://github.com/francescotonini/multimodal-across-domains-gaze-target-detection. Below, we split our proposal into two sections: Multimodal Network (Sec. 3.1) and Domain Adaption (Sec. 3.2), and describe each aforementioned module in detail. Lastly, Sec. 3.3 supplies the implementation details.

3.1 Multimodal Network

Head Network. Given the RGB scene image S_i, we crop the head of the person-of-interest to obtained head image H_i. H_i is processed by the head network's backbone that maps the original representation H_i into a feature embedding $\mathbf{e}_i^H$. Such features are average pooled and processed by a set of linear layers that outputs an attention map. The outcome of the attention map $\mathbf{attn}_i^H$ is multiplied by the scene and depth feature embeddings.

Scene Network. The scene network shares the backbone structure of the head and depth networks. This module takes as the input the concatenation of the RGB scene image S_i and the binary head location mask M_i that encodes the position of the person's head in the scene image. Each channel of the feature embedding of the scene network $\mathbf{e}_i^S$ is multiplied by the attention map $\mathbf{attn}_i^H$ generated by the head network:

$$\mathbf{e}_i^{S*} = \mathbf{e}_i^S \otimes \mathbf{attn}_i^H, \tag{1}$$

where $\otimes$ is the channel-wise multiplication. By multiplying the output of the scene network's backbone with the attention map, we force the network to focus on objects in the scene that are relevant with respect to the person-of-interest and its head orientation. This is in line with SOTA [5, 10].

Depth Network. The depth network shares the same backbone structure and the input shape of the scene network **SN**. This module takes as input the depth map D_i of the scene and the binary head location mask M_i. The feature embeddings $\mathbf{e}_i^D$ from the depth backbone are multiplied by the head attention map $\mathbf{attn}_i^H$:

$$\mathbf{e}_i^{D*} = \mathbf{e}_i^D \otimes \mathbf{attn}_i^H, \tag{2}$$

where $\otimes$ is the channel-wise multiplication.

Fusion & Prediction Network. The feature embeddings from the head network $\mathbf{e}_i^H$, the *attended* scene $\mathbf{e}_i^{S*}$ and the *attended* depth embeddings $\mathbf{e}_i^{D*}$ are the inputs of the fusion and prediction network **F**. The fusion module contains two encoder **ES** and **ED** that process

the concatenation of $\mathbf{e}_i^H$, $\mathbf{e}_i^{S*}$, and $\mathbf{e}_i^H$, $\mathbf{e}_i^{D*}$ independently:

$$\mathbf{e}_i^{HS*} = \mathbf{ES}(concat(\mathbf{e}_i^H, \mathbf{e}_i^{S*})), \tag{3}$$

$$\mathbf{e}_i^{HD*} = \mathbf{ED}(concat(\mathbf{e}_i^H, \mathbf{e}_i^{D*})), \tag{4}$$

where $concat(\mathbf{A}, \mathbf{B})$ is the channel-wise concatenation of the feature embeddings of $\mathbf{A}$ and $\mathbf{B}$.

The prediction module outputs the 2D gaze heatmap $\mathbf{Heat}_i$ and the in/out of frame probability $\mathbf{InOut}_i$. To obtain the 2D gaze heatmap $\mathbf{Heat}_i$, this module uses a multi-layer decoder $\mathbf{D}$ that takes as input the channel-wise summation $\oplus$ of scene and depth embeddings:

$$\mathbf{Heat}_i = \mathbf{D}(\mathbf{e}_i^{HS*} \oplus \mathbf{e}_i^{HD*}), \tag{5}$$

Furthermore, the channel-wise outer product [39] between scene and depth embeddings is input to a smaller encoder **EInOut** that produces the $\mathbf{InOut}_i$:

$$\mathbf{InOut}_i = \mathbf{EInOut}(outer(\mathbf{e}_i^{HS*}, \mathbf{e}_i^{HD*})), \tag{6}$$

3.2 Domain Adaptation

The multimodal network is enriched by multiple domain adaptation (DA) components that improve the performance of our proposed method when it is trained and tested across datasets. More specifically, our DA components attempt to improve the performance of our three networks: head, scene, and depth.

First, we introduce a domain classifier that performs a binary classification between the source and target domain represented in terms of the embeddings of the head backbone $\mathbf{e}_i^H$. The domain classifier is connected to head backbone via Gradient Reversal Layer (GRL) [13], which multiplies the gradient by a certain negative constant during the backpropagation-based training. In other words, the domain classification loss is minimized for both source and target samples and the GRL ensures that the learned features are as indistinguishable as possible [13] (we simply show this as L_{GRL} in Fig. 1). The choice of adapting GRL is due to the fact that it has been one of the most popular UDA method, adapted to handle domain-shift problem for several applications, e.g., intention detection [61], view-invariant action recognition [33], object recognition [42] and re-identification [13].

Second, two additional decoders are integrated to the proposed method to reconstruct the original input $\mathbf{S}_i$ using the embedding of the depth network $\mathbf{e}_i^D$, and vice-versa. Thus, we perform a modality translation from RGB to depth, and depth to RGB. The implementation details clarify the associated loss functions.

3.3 Implementation Details

We implemented our model in PyTorch. The input to scene, depth, and head networks are normalized and resized to 224×224. The backbones are based on ResNet-50 with an additional layer that creates a feature embedding of 1024 channels of size 7×7 in line with [4, 5]. The scene backbone is pre-trained on Places dataset [58], and the head backbone is pre-trained on Eyediap dataset [12] as applied in [4, 5, 10, 18], which show improved performance compared to using Vanilla ResNet-50. The depth backbone is also pre-trained on Places dataset [58], after obtaining magma-colored depth maps by applying [34] for each RGB image in the training set of [58].

To perform fair analysis with the prior art [4, 5, 10, 23, 35, 36], the ground-truth gaze heatmap is obtained by plotting a Gaussian distribution around the center of gaze, i.e., the ground-truth gaze coordinate with respect to the scene image. The total loss of **our multimodal network** is a weighted sum of the Mean Squared Error (*MSE*) loss on the gaze heatmap $L_{heatmap}$ and a binary cross entropy (*BCE*) loss $L_{in/out}$ for the in/out output:

$$L_{total} = w_{heatmap}\, L_{heatmap} + w_{in/out}\, L_{in/out}, \tag{7}$$

where $w_{heatmap}$ and $w_{in/out}$ are the learnable weights.

Our multimodal network was trained from scratch on GazeFollow [35] for 70 epochs with the batch size of 16 and the learning rate of 2.5×10^{-4}. Afterwards, we fine-tuned the model on VideoAttentionTarget [5] following the implementation of SOTA [5, 10, 41]. Furthermore, we trained our multimodal network from scratch on GOO [41] for 70 epochs with the batch size of 16 and the learning rate of 2.5×10^{-4}.

To perform proposed **domain adaptation** method, at each step of the training, we forward a batch from the source domain followed by a batch from the target domain. The training was for up to 70 epochs, with a batch size of 16 and the learning rate of 2.5×10^{-4}. Due to the absence of the *in/out* annotation on multiple datasets (see Sec. 5 for more details), while applying DA, we do not minimize the cross entropy loss: $L_{in/out}$. The total loss ($L_{total_{w/DA}}$) while applying our DA module includes three additional losses: the cross entropy loss on the head domain classifier (L_{GRL}), the *MSE* reconstruction losses from RGB to Depth ($L_{RGB \to Depth}$) and Depth to RGB ($L_{Depth \to RGB}$), shown as:

$$\begin{aligned} L_{total_{w/DA}} &= w_{heatmap}\, L_{heatmap} + w_{GRL}\, L_{GRL} \\ &+ w_{RGB \to Depth}\, L_{RGB \to Depth} + w_{Depth \to RGB}\, L_{Depth \to RGB}, \end{aligned} \tag{8}$$

where $w_{heatmap}$, w_{GRL}, $w_{RGB \to Depth}$, and $w_{Depth \to RGB}$ are the learnable weights.

4 EXPERIMENTAL ANALYSIS

We conducted a comprehensive analysis to evaluate the performance of our method. Sec. 4.2 presents an ablation study for our multimodal network to show the contribution of the scene, depth and head networks. It also includes an extensive investigation regarding modality fusion. Sec. 4.3 compares the performance of the proposed multimodal network against to the prior art. In Sec. 4.4, we study gaze target detection task across datasets. The experiments given in Sec. 4.2- 4.4 do not apply any DA method, while the experiments in Sec. 4.5 corresponds to applying DA. Our method (with / without DA) achieves the state-of-the-art results on all datasets in all experiments. Finally, we present the qualitative results of the proposed method (with / without DA) in Sec. 4.6.

4.1 Datasets & Evaluation Metrics

The proposed method is evaluated on three benchmark datasets: GazeFollow [35], VideoAttentionTarget [5] and Gaze On Objects (GOO) [41]. We follow the standard training and test splits of each dataset to supply fair comparisons with SOTA.

Datasets. *GazeFollow* [35] dataset includes more than 120K images from various classification and detection datasets (i.e., SUN [50], COCO [24], Actions-40 [54], PASCAL [9], and Places [58]), with more than 130K annotations of head locations and the corresponding gaze points. *VideoAttentionTarget* [5] is a collection of 1331 video clips from various sources on YouTube. The annotations include more than 160K frame-level head bounding boxes and 110K gaze targets inside the scene. *Gaze On Objects (GOO)* [41] dataset is a collection of images of shelves with 24 classes of groceries. In each image, a person looks at one object on a shelf. Objects in the scene are annotated with their bounding box and class. GOO is the first dataset in the gaze target detection task that uses both real and synthetic data. Out of the 200K images of the datasets, 8K are captured from a real environment, while 192K are generated using a 3D engine that reconstructs the real environment. In this paper, we use the images belonging to real environment.

Evaluation Metrics. The following metrics were adopted to evaluate the performance of the proposed model in line with the SOTA [4, 5, 10, 23, 35]. *Heatmap Area Under Curve (AUC %)* [19] is to asses the confidence of the predicted heatmap with respect to the ground-truth. *Average distance (Avg.Dist.)* stands for the Euclidean distance between the predicted gaze location and the ground-truth gaze point.

4.2 Ablation Study & Modality Fusion

To better investigate the contribution of different components of our model (i.e., scene, head and depth networks) and to compare the effectiveness of the different modality fusion techniques, we trained the following variations. **1) Scene network only:** We remove the head network **HN** and the depth network **DN**. The head features concatenated with the scene features are not provided, thus the only way to make the attention map $\mathbf{attn}_i^H$ is through the head location mask $\mathbf{M}_i$. **2) Scene and head networks:** This stands for removing the depth network **DN** while the scene **SN** and the head networks **HN** remain. The input of the scene network is the concatenation of the scene image $\mathbf{S}_i$ and the head location mask $\mathbf{M}_i$. **3) Grayscale depth and head networks:** This refers to removing the scene network **SN** while keeping the head **HN** and the depth networks **DN**. The input of the depth network is the concatenation of the *grayscale* depth map $\mathbf{D}_i$ and the head location mask $\mathbf{M}_i$. **4) Colored depth and head networks:** This is a similar set up with (3) when the depth map $\mathbf{D}_i$ is colored by magma-colormap (as shown in Fig. 1). **5) Early fusion V1:** We replace the head location mask $\mathbf{M}_i$ with the *grayscale* depth map $\mathbf{D}_i$ while the depth network **DN** is completely removed. In other words, this setup includes head network **HN** and the scene network **SN** when the input tensor of the scene network has four channels (three for scene image $\mathbf{S}_i$ and one for *grayscale* depth map $\mathbf{D}_i$). **6) Early fusion V2:** This refers to removing the depth network **DN** and feeding **SN** with the input tensors having five channels (three for scene image $\mathbf{S}_i$, one for head location mask $\mathbf{M}_i$, and one for *grayscale* depth map $\mathbf{D}_i$). **7) Early fusion V3:** This is a similar set up to (6) when the input tensors have seven channels (three for scene image $\mathbf{S}_i$, one for head location mask $\mathbf{M}_i$ and three for colored depth map $\mathbf{D}_i$). **8) Early fusion V4:** This refers to having a single encoder network instead of having **ES** and **ED** whose inputs are the concatenation of the scene feature map, depth feature map and the head feature map. **9) Depth-Aware Scene Convolutional Network (Early Fusion V5):** We remove the proposed **DN** and replace the **SN** with the depth-aware scene convolutional network of [46] when the inputs of it are the grayscale depth map and the RGB scene images in parallel. **10) Late fusion with concatenation:** We replace the proposed summation operation in Eq. 5 with concatenation, which is applied before the multi-layer decoder **D**. **11) Late fusion with outer product:** It refers to replacing the applied summation operation in Eq. 5 with the channel-wise outer product proposed in [39]. These aforementioned variations were tested on the GazeFollow [35] dataset. The most competitive ones were also validated on the VideoAttentionTarget [5] dataset. The corresponding results are given in Table 1.

Overall, the results show that all components of the proposed multimodal network are important to achieve the best performance. The experiments 1, 2, 3 and 4 allow us to understand the contribution of the head, depth and the scene networks, respectively. Out of three, the most crucial component is the head network (improving the AUC by 16.9% and decreasing the Avg.Dist. by 0.117 in GazeFollow dataset [35]), which gives intuition regarding the head orientation of a person in the scene, and allowing scene and the depth networks to pay more attention to the features that are more likely to be attended to. The second most contributing component is the scene network (performs up to +4.6% AUC and -0.072 Avg.Dist., compared to depth+head network in GazeFollow dataset [35]). Still the usage of scene network without the depth network fails to detect the target in a different depth level than the person who is gazing and, indeed scene+depth+head (proposed method) achieves +0.6% AUC and -0.002 Avg.Dist. compared to scene+head in GazeFollow dataset [35] while the performance improvement of the proposed method is higher in VideoAttentionTarget [5] (+3.4% AUC and -0.01 Avg.Dist.). On the other hand, when the depth and head networks remain and the scene network is removed, (experiments 3 and 4), one can be in favor of using the colored depth map instead of grayscale depth map (+1.8% AUC and -0.05 Avg.Dist. in GazeFollow dataset [35]).

The examination regarding how to combine the modalities is composed of applying various early fusion and late fusion experiments. The early fusion experiments (Exp. 5, 6 and 7) combine the modalities in the input space. Discarding head location mask (Exp. 5) decreases the performance compared to Exp. 6 and 7 by up to -1.2% AUC and +0.008 Avg.Dist. in GazeFollow dataset [35]. In early fusion setups, using grayscale depth map surpasses the colored version by +0.9% AUC and -0.002 Avg.Dist. in GazeFollow dataset [35]. However, this trend was not observed for VideoAttentionTarget [5], in which using colored depth map achieved slightly better results than grayscale depth map in terms of AUC (+0.3%). It is important to notice that these experiments are relatively lightweight as having only scene convolutional network compared to the proposed method (late fusion) and Exp. 8. Exp. 8 differs from Exp. 5, 6 and 7 as it combines the embeddings of the modalities before being encoded. It performs worse than other early fusion variations by up to -1.1% AUC in GazeFollow dataset [35], instead performed better than others by up to +0.8% AUC and -0.007 Avg.Dist. in

Table 1: Results of ablation study and the modality fusion on the GazeFollow [35] and VideoAttentionTarget [5] datasets. Exp. follows the order of variations (1-11) described in Sec.4.2. For ease of reading, the ablation study, early fusion and late fusion are given in yellow, pink, and blue, respectively. The best results (the higher the AUC and the lower the Avg.Dist.) are shown in bold.

Exp.	1	2	3	4	5	6	7	8	9	10	11	**Ours**
GazeFollow [35]												
AUC (%)	75.8	92.1	87.5	89.3	91.2	92.4	91.5	91.3	90.0	92.6	92.6	**92.7**
Avg.Dist.	0.258	0.143	0.215	0.166	0.157	0.149	0.151	0.143	0.183	0.143	0.142	**0.141**
VideoAttentionTarget [5]												
AUC (%)	-	90.6	-	-	-	93.0	93.3	93.8	-	90.5	90.6	**94.0**
Avg.Dist.	-	0.139	-	-	-	**0.129**	0.139	0.132	-	0.143	0.141	**0.129**

VideoAttentionTarget [5]. We also adapted the recent work Depth-Aware Scene Convolutional Network [46] by replacing it with the scene convolutional network. However, this lowered the results up to -2.4% for AUC and +0.04 for Avg.Dist. compared to other early fusion applications.

We further investigate different ways of applying late fusion. In GazeFollow [35] dataset, we observed that different late fusion operations (i.e., concatenation, channel-wise outer product [39] and summation in Eq. 5) perform on par while the proposed summation operation achieves slightly better AUC (+0.1%) than others. It is important to notice that, for that dataset, all early fusion methods (Exp. 5-9) achieve worse results compared to the all late fusion methods (Exp. 10-11) with the drop in the margin of 0.2-3.3% for AUC and the increase in the margin of 0.007-0.024 for Avg.Dist. On the other hand, the proposed late fusion (see Eq. 5) achieves remarkable results on VideoAttentionTarget [5] compared to other late fusion methods. That is up to +3.5% for AUC and -0.014 for Avg.Dist. This also presents that the effectiveness of the proposed method generalizes better across different datasets compared to the other early and late fusion approaches tested.

4.3 Comparisons with the State-of-the-art

We compare our multimodal network with several SOTA in Table 2. These comparisons include the standard gaze analysis baselines, namely: *i)* random, *ii)* center bias, and *iii)* fixed bias, whose results are taken from [35]. Random stands for generating a heatmap per pixel by sampling the values from a Gaussian distribution. In center bias, the prediction is always the center of the image. In fixed bias, the location of the prediction is in terms of the average of fixations from the training set for the heads located to a similar area with the test image.

Our method achieves better results compared to all counterparts, and becomes SOTA for all datasets in terms of AUC. It surpasses even the human performance in GazeFollow [35] (+0.3% AUC) and VideoAttentionTarget [5] (+1.9% AUC) datasets. In particular, its relative performance improvements in VideoAttentionTarget [5] and GOO [41] datasets are obtrusive (3.5-11% and 1.8-12.2% AUC, respectively). In terms of Avg.Dist., our method falls behind Fang et al. [10] and Jin et al. [18] while performing better than others. It is important to notice that Fang et al. [10] presents more complex and less lightweight model compared to ours by having additional components to *i)* extract the head pose, *ii)* detect the eyes and *iii)* extract the eye features, which might be infeasible to perform correctly in real-life application. On the other hand, we can argue that the auxilary networks used in Jin et al. [18] for gathering the 3D-gaze orientation information and depth map help to improve Avg.Dist., while that method performs poorer than ours and many other SOTA in terms of AUC.

Moreover, one can observe that, even though being a spatial model, our method outperforms Chong et al. [5], which includes Convolutional LSTM network. Consequently, we can draw a conclusion that integrating the depth map generated from the RGB scene images, i.e., a multimodal approach such as ours, Fang et al. [10] or Jin et al. [18], results in better gaze target detection performance compared to relying on RGB videos, i.e., performing spatio-temporal data processing as in [5]. The late fusion results obtained by slightly tuning the proposed method (see Table 1, Exp. 10 and 11) also confirm this conclusion by outperforming all the prior art. Besides, it is important to notice that some of the variations presented in Sec. 4.2, while being less effective than the proposed method, are still able to surpass the existing methods [4, 5, 10, 18, 23, 35] particularly when tested on VideoAttentionTarget [5] dataset.

4.4 Gaze Target Detection Across Datasets

This section examines the domain-shift problem for gaze target detection task in images and videos. To do so, we trained the model of Chong et al. [5], and the proposed method on one dataset while the trained models were tested on a completely different dataset. The corresponding results are given in Table 3. As seen, our method outperforms Chong et al. [5] in all cross-domain analysis. However, it is important to notice that both method (even though they were trained on in-the-wild datasets) suffer a lot (up to -36.5% AUC and +0.3% Avg.Dist.) when they were trained and tested on different domains with respect to training and testing them on the same domain. Consequently, we can argue that addressing domain-shift problem for gaze target detection task is inevitable.

4.5 Domain Adaptation Results

Table 4 shows the results of the proposed domain adaption method (see Sec. 3.2). We evaluate our results as compared to Table 3. Additionally, we adapted the SOTA domain adaptation method Ferreri et al. [11] to make comparisons among it's performance and ours. The corresponding results as well as an ablation study for the proposed DA module, are involved into the Table 4.

Table 2: Evaluation on benchmark datasets. The best results, the higher the *AUC* and the lower the average distance (*Avg.Dist.*) is better, are shown in bold. ⋆ indicates our training. ⋄ taken from [41]. *Ours* refers to the proposed multimodal network without domain adaptation. See text for the description of Random, Center and Fixed Bias.

	GazeFollow [35]		VideoAttentionTarget [5]		GOO [41]	
	AUC	Avg.Dist.	AUC	Avg.Dist.	AUC	Avg.Dist.
Random	50.4	0.484	50.5	0.458	-	-
Center	63.3	0.313	-	-	-	-
Fixed Bias	67.4	0.306	72.8	0.326	-	-
Recasens et al. (2015) [35]	87.8	0.190	-	-	85.0⋄	0.220⋄
Chong et al. (2018) [4]	89.6	0.187	83.0	0.193	-	-
Lian et al. (2018) [23]	90.6	0.145	-	-	84.0⋄	0.321⋄
Chong et al. (2020) [5]	92.1	0.137	86.0	0.134	79.6⋄	0.252⋄
Chong et al. (2020) [5]⋆	92.2	0.143	86.2	0.136	90.0	0.190
Fang et al. (2021) [10]	92.2	0.124	90.5	**0.108**	-	-
Jin et al. (2022) [18]	92.0	**0.118**	90.1	0.116	-	-
Ours	**92.7**	0.141	**94.0**	0.129	**91.8**	**0.164**
Human Performance	92.4	0.096	92.1	0.051	-	-

Table 3: Evaluation of gaze target detection performance across datasets. The drop with respect to the same-dataset evaluation are given in parenthesis. Results in black are better than its counterpart. ⋆ indicates our training.

	Trained on	Tested on	AUC (%)	Avg.Dist.
Chong et al. (2020) [5]⋆	GazeFollow	GOO	77.3 (13.3↓)	**0.270** (0.1↓)
Ours	GazeFollow	GOO	**78.3** (13.5↓)	0.284 (0.1↓)
Chong et al. (2020) [5]⋆	VideoAttentionTarget	GOO	68.2 (22.4↓)	0.311 (0.1↓)
Ours	VideoAttentionTarget	GOO	**69.1** (22.7↓)	**0.274** (0.1↓)
Chong et al. (2020) [5]⋆	GOO	GazeFollow	62.5 (29.6↓)	0.410 (0.3↓)
Ours	GOO	GazeFollow	**62.9** (29.8↓)	**0.401** (0.3↓)
Chong et al. (2020) [5]⋆	GOO	VideoAttentionTarget	55.1 (30.9↓)	0.458 (0.3↓)
Ours	GOO	VideoAttentionTarget	**57.5** (36.5↓)	**0.446** (0.3↓)

Ferreri et al. (2021) [11] is a SOTA for multimodal RGB-D scene recognition task, which takes as the inputs scene images in the format of RGB and depth. That pipeline [11] matches with our multimodal network. Consequently, we adapt Ferreri et al. [11] into our multimodal network, also noticing that there is no DA method proposed for gaze target detection that we can conduct a comparison. To do so, we added two additional decoders to the proposed multimodal network to perform modality translation from RGB to depth, and depth to RGB. These decoders attempt to reconstruct the original input (e.g., scene image) using features extracted from the other modality (e.g., depth embeddings). Such images were then compared against to the original input. An additional frozen ResNet-18 was also integrated to the proposed multimodal network to perform content similarity loss between reconstructed images and the original images. Such loss is obtained by calculating the $L1$ loss of multiple layers of the additional ResNet-18 network [11]. We trained our multimodal network using the code of [11] with a learning rate of 2.5×10^{-4} and for up to 40 epochs. We noticed that the test performances in earlier epochs (e.g., 3, 5) are better in all settings. We report the best of all test results of Ferreri et al. [11] in Table 4. Our DA pipeline differs from Ferreri et al. [11] in terms of the following aspects: *i)* instead of content similarity loss, we rely on a reconstruction loss between reconstructed and original images, and *ii)* we also use a smaller decoder that shares the same structure of the *fusion & prediction network*'s decoder. Precisely, it is composed of four *convolutional + ReLU* blocks that reconstruct the RGB / Depth image starting from embeddings of Depth / RGB backbones. Compared to Ferreri et al. [11], one can notice that we present a much simpler and much lightweight DA component, which performs between *scene* and *depth* images. Additionally, integrating GRL further powers up the *head features* consistency across source and target datasets. Indeed, the ablation study in Table 4 (applied when the source dataset is VideoAttentionTarget and target dataset is GOO) proves that each loss function of our DA module is important, and using them altogether results in the best performance.

The proposed DA notably improves the performance: AUC (1.3-15.6% more) and Avg.Dist. (0.017-0.086 less) compared to without applying DA. These improvements are higher than Ferreri et al. [11] in all settings in terms of AUC. In terms of Avg.Dist., for GOO → VideoAttentionTarget, [11] performs better than the proposed DA while for the rest of the settings the proposed DA outperforms Ferreri et al. [11]. Also, one can observe that the performance across domains does not always raise up by applying Ferreri et al. [11] (e.g., GazeFollow → GOO and vice-versa).

Table 4: Evaluation of domain adaptation methods for gaze target detection. The performance improvement (↑) or drop (↓) with respect to the *Ours* in Table 3 are given in parenthesis. Notice that higher values of AUC and lower values of Avg.Dist. mean an improvement. Results in black are better than its counterpart. Full refers to Eq. 8.

	Source	Target	AUC (%)	Avg.Dist.
Ferreri et al. (2021) [11]	GazeFollow	GOO	66.1 (12.2↓)	0.327 (0.043↓)
Ours (Full)	GazeFollow	GOO	**84.0** (5.7↑)	**0.238** (0.046↑)
Ferreri et al. (2021) [11]	VideoAttentionTarget	GOO	61.8 (7.3↓)	0.388 (0.114↓)
Ours (L_{GRL})	VideoAttentionTarget	GOO	75.3 (6.2↑)	0.300 (0.026↓)
Ours ($L_{RGB \to Depth}$ + $L_{Depth \to RGB}$)	VideoAttentionTarget	GOO	69.9 (0.8↑)	0.301 (0.027↓)
Ours (Full)	VideoAttentionTarget	GOO	**77.5** (8.4↑)	**0.257** (0.017↑)
Ferreri et al. (2021) [11]	GOO	GazeFollow	62.5 (0.4↓)	**0.412** (0.011↓)
Ours (Full)	GOO	GazeFollow	**64.2** (1.3↑)	0.413 (0.012↓)
Ferreri et al. (2021) [11]	GOO	VideoAttentionTarget	69.2 (11.7↑)	**0.325** (0.121↑)
Ours (Full)	GOO	VideoAttentionTarget	**73.1** (15.6↑)	0.360 (0.086↑)

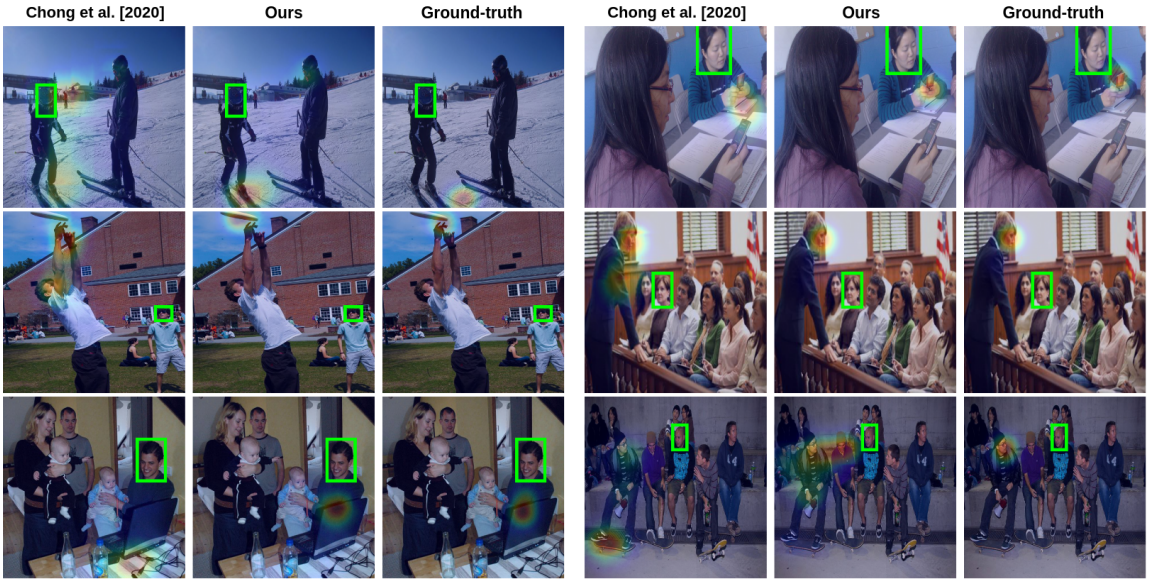

Figure 2: Qualitative results on the GazeFollow [35] dataset. We evaluate the performance of our multimodal network, Chong et al. [5] with respect to the ground-truth data. Green bounding boxes are taken from the corresponding dataset, referring to the cropped head image of the person whose gaze target to be detected.

4.6 Qualitative Results

Fig. 2 shows qualitative results on GazeFollow [35] dataset, in which we present our multimodal network's (i.e., no DA) and Chong et al. [5]'s results together with the ground-truth. As seen, our neural network is able to capture gaze in challenging and dynamic scenes as well as producing more compact heatmaps (implying less Avg.Dist.) compared to Chong et al. [5]. In Fig. 3, we demonstrate results of our model with and without domain adaptation. The corresponding images are obtained when the source dataset is GazeFollow [35] and the target dataset is GOO [41]. Given the ground-truth data, one can observe that our domain adaptation method notably improves the gaze target detection results compared to our multimodal network without domain adaptation.

5 DISCUSSION

We have presented a novel multimodal deep architecture in order to identify where the person, in an image taken from the third perspective, is looking. Our spatial model is composed of three-pathways simultaneously processing *i)* the head image belonging to the person whose gaze target is to be detected (i.e., person-of-interest), *ii)* the scene image and *iii)* the depth maps, both supplying the context information. It is distinguishable from the prior art as it does not rely on supervision of gaze angle, does not require explicit head orientation information or the location of the eyes of the person-of-interest. Extensive quantitative and qualitative evaluations demonstrate that the proposed method performs favorably against the existing approaches. Additionally, our investigations regarding joint-learning of multiple modalities resulted in several

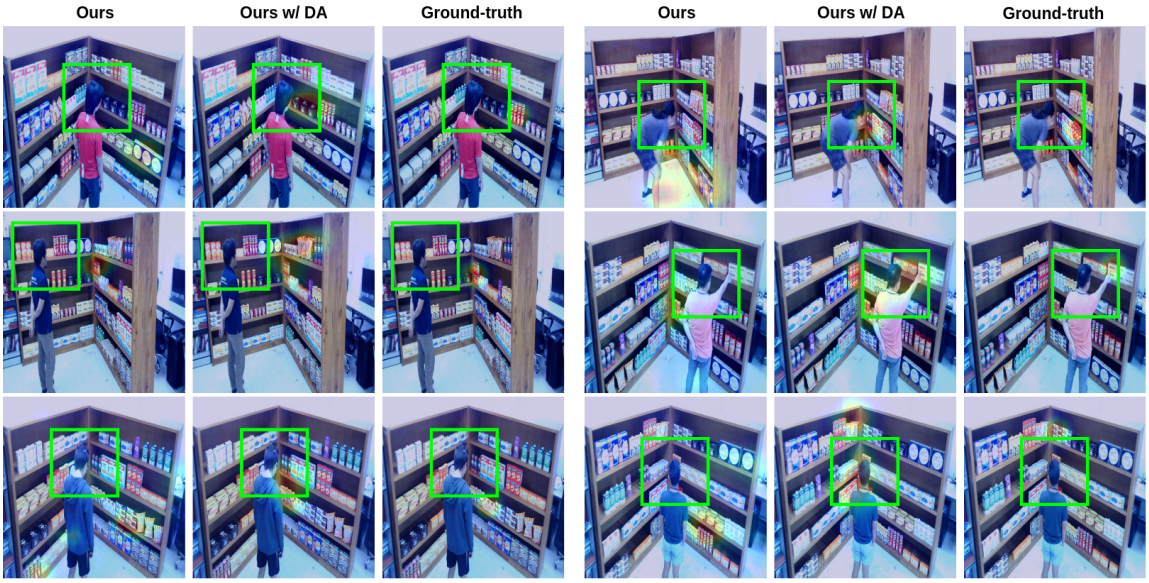

Figure 3: Qualitative results of our proposal with and without the domain adaptation component when the source data is GazeFollow [35] and the target data is GOO [41]. Green bounding boxes are taken from the corresponding dataset, referring to the cropped head image of the person whose gaze target to be detected.

variations of the proposed method. Some of these variations, notice that they have never presented in an earlier work, also outperforms several SOTA.

First time in this paper, we also studied the domain adaption (DA) for gaze target detection. To do so, we have injected new DA components to the described multimodal network. Our proposal enhanced the performance on target datasets as well as performing better than the DA SOTA. It is important to mention that the used datasets were all collected in unconstrained situations, including complex human-human social interactions and/or human-object interactions. The effective results of the proposed method on these benchmarks, and particularly its capability to handle domain-shift problem, potentially makes it stronger than the counterparts when it is integrated to real-life applications. Inline with SOTA [5, 10], the proposed method not only detects the gaze targets located in the scene but also able to declare if the gaze target is out-of-the-scene. In this paper, we have not evaluated our method in terms of out-of-frame precision [5, 10] (consequently, it is discarded from the contributions as well). This is due to lack of corresponding annotations in multiple datasets, which does not allow us to train and/or test our pipeline, particularly the DA component. One potential future work will be supplying out-of-frame annotations for all existing benchmarks (at least for their test splits). Additionally, the proposed method will be tested on unconstrained human-robot interaction scenarios targeting assistive robotics application in hospitals using the real-life dataset collected by the EU Horizon 2020 SPRING project (No. 871245). We will also investigate integrating transformers into our multimodal pipeline to better exploit the head attention over the scene and depth networks.

ACKNOWLEDGMENTS

This work was supported by the EU Horizon 2020 SPRING project (No. 871245).

REFERENCES

[1] Cigdem Beyan, Nicolò Carissimi, Francesca Capozzi, Sebastiano Vascon, Matteo Bustreo, Antonio Pierro, Cristina Becchio, and Vittorio Murino. 2016. Detecting Emergent Leader in a Meeting Environment Using Nonverbal Visual Features Only. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 317–324. https://doi.org/10.1145/2993148.2993175

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. 2017. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3722–3731.

[3] Ernesto Brau, Jinyan Guan, Tanya Jeffries, and Kobus Barnard. 2018. Multiple-gaze geometry: Inferring novel 3d locations from gazes observed in monocular video. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 612–630.

[4] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. 2018. Connecting Gaze, Scene, and Attention: Generalized Attention Estimation via Joint Modeling of Gaze and Scene Saliency. In *The European Conference on Computer Vision (ECCV)*.

[5] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. 2020. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5396–5406.

[6] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Chi Su, Qingming Huang, and Qi Tian. 2020. Gradually vanishing bridge for adversarial domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12455–12464.

[7] Victor G. Turrisi da Costa, Giacomo Zara, Paolo Rota, Thiago Oliveira-Santos, Nicu Sebe, Vittorio Murino, and Elisa Ricci. 2022. Dual-Head Contrastive Domain Adaptation for Video Action Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1181–1190.

[8] Murtada Dohan and Mu Mu. 2019. Understanding User Attention In VR Using Gaze Controlled Games. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video* (Salford (Manchester), United Kingdom) *(TVX '19)*. Association for Computing Machinery, New York, NY, USA, 167–173. https://doi.org/10.1145/3317697.3325118

[9] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vision* 88, 2 (jun 2010), 303–338. https://doi.org/10.1007/s11263-009-0275-4

[10] Yi Fang, Jiapeng Tang, Wang Shen, Wei Shen, Xiao Gu, Li Song, and Guangtao Zhai. 2021. Dual attention guided gaze target detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11390–11399.

[11] Andrea Ferreri, Silvia Bucci, and Tatiana Tommasi. 2021. Multi-Modal RGB-D Scene Recognition Across Domains. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2199–2208.

[12] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 2014. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proceedings of the symposium on eye tracking research and applications*. 255–258.

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[14] Zidong Guo, Zejian Yuan, Chong Zhang, Wanchao Chi, Yonggen Ling, and Shenghao Zhang. 2020. Domain adaptation gaze estimation by embedding with prediction consistency. In *Proceedings of the Asian Conference on Computer Vision*.

[15] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. 2016. Learning With Side Information Through Modality Hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Judy Hoffman, Saurabh Gupta, Jian Leong, Sergio Guadarrama, and Trevor Darrell. 2016. Cross-modal adaptation for RGB-D detection. In *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 5032–5039.

[17] Zhengxi Hu, Dingye Yang, Shilei Cheng, Lei Zhou, Shichao Wu, and Jingtai Liu. 2022. We Know Where They Are Looking at From the RGB-D Camera: Gaze Following in 3D. *IEEE Transactions on Instrumentation and Measurement* 71 (2022), 1–14. https://doi.org/10.1109/TIM.2022.3160534

[18] Tianlei Jin, Qizhi Yu, Shiqiang Zhu, Zheyuan Lin, Jie Ren, Yuanhai Zhou, and Wei Song. 2022. Depth-aware gaze-following via auxiliary networks for robotics. *Engineering Applications of Artificial Intelligence* 113 (2022), 104924.

[19] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

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

[21] Xiao Li, Min Fang, Ju-Jie Zhang, and Jinqiao Wu. 2017. Domain Adaptation from RGB-D to RGB Images. *Signal Process.* 131, C (feb 2017), 27–35. https://doi.org/10.1016/j.sigpro.2016.07.018

[22] Yin Li, Miao Liu, and Jame Rehg. 2021. In the eye of the beholder: Gaze and actions in first person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[23] Dongze Lian, Zehao Yu, and Shenghua Gao. 2018. Believe it or not, we know what you are looking at!. In *Asian Conference on Computer Vision*. Springer, 35–50.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[25] Meng Liu, Youfu Li, and Hai Liu. 2020. 3D Gaze Estimation for Head-Mounted Eye Tracking System With Auto-Calibration Method. *IEEE Access* 8 (2020), 104207–104215. https://doi.org/10.1109/ACCESS.2020.2999633

[26] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. PMLR, 97–105.

[27] Manuel J. Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. 2019. LAEO-Net: Revisiting People Looking at Each Other in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Benoît Massé, Silèye Ba, and Radu Horaud. 2017. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE transactions on pattern analysis and machine intelligence* 40, 11 (2017), 2711–2724.

[29] Benoit Massé, Stéphane Lathuilière, Pablo Mesejo, and Radu Horaud. 2019. Extended gaze following: Detecting objects in videos beyond the camera field of view. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 1–8.

[30] Kyle Min and Jason J. Corso. 2021. Integrating Human Gaze Into Attention for Egocentric Activity Recognition. In *Proc. of IEEE/CVF WACV*. 1069–1078.

[31] Philipp Müller, Ekta Sood, and Andreas Bulling. 2020. Anticipating averted gaze in dyadic interactions. In *ACM Symposium on Eye Tracking Research and Applications*. 1–10.

[32] Radoslaw Niewiadomski, Lea Chauvigne, Maurizio Mancini, and Antonio Camurri. 2018. Towards a Model of Nonverbal Leadership in Unstructured Joint Physical Activity. In *Proc. of MoCo* (Genoa, Italy) *(MOCO '18)*. Association for Computing Machinery, Article 20, 8 pages. https://doi.org/10.1145/3212721.3212816

[33] Giancarlo Paoletti, Jacopo Cavazza, Cigdem Beyan, and Alessio Del Bue. 2021. Unsupervised Human Action Recognition with Skeletal Graph Laplacian and Self-Supervised Viewpoints Invariance. In *The 32nd British Machine Vision Conference (BMVC)*.

[34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[35] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking?. In *Advances in Neural Information Processing Systems*, Vol. 28. Curran Associates, Inc.

[36] Adrià Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. 2017. Following Gaze in Video. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 1444–1452. https://doi.org/10.1109/ICCV.2017.160

[37] Boris Schauerte and Rainer Stiefelhagen. 2014. "Look at this!" learning to guide visual saliency in human-robot interaction. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 995–1002. https://doi.org/10.1109/IROS.2014.6942680

[38] Luciano Spinello and Kai O Arras. 2012. Leveraging rgb-d data: Adaptive fusion and domain adaptation for object detection. In *2012 IEEE International Conference on Robotics and Automation*. IEEE, 4469–4474.

[39] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8992–8999.

[40] Sanket Kumar Thakur, Cigdem Beyan, Pietro Morerio, and Alessio Del Bue. 2021. Predicting Gaze from Egocentric Social Interaction Videos and IMU Data. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 717–722.

[41] Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. 2021. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3125–3133.

[42] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. 2015. Simultaneous deep transfer across domains and tasks. In *Proceedings of the IEEE international conference on computer vision*. 4068–4076.

[43] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial Discriminative Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[44] Jing Wang and Kuangen Zhang. 2019. Unsupervised domain adaptation learning algorithm for rgb-d staircase recognition. *arXiv preprint arXiv:1903.01212* (2019).

[45] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. 2021. Domain Adaptive Semantic Segmentation With Self-Supervised Depth Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 8515–8525.

[46] Weiyue Wang and Ulrich Neumann. 2018. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 135–150.

[47] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. 2018. Where and Why are They Looking? Jointly Inferring Human Attention and Intentions in Complex Tasks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6801–6809. https://doi.org/10.1109/CVPR.2018.00711

[48] Garrett Wilson and Diane J. Cook. 2020. A Survey of Unsupervised Deep Domain Adaptation. *ACM Trans. Intell. Syst. Technol.* 11, 5, Article 51 (jul 2020), 46 pages. https://doi.org/10.1145/3400066

[49] Yi-Leh Wu, Chun-Tsai Yeh, Wei-Chih Hung, and Cheng-Yuan Tang. 2014. Gaze direction estimation using support vector machine with active appearance model. *Multimedia Tools Appl.* 70 (2014), 2037–2062. https://doi.org/10.1007/s11042-012-1220-z

[50] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 3485–3492.

[51] Jiaolong Xu, Liang Xiao, and Antonio M López. 2019. Self-supervised domain adaptation for computer vision tasks. *IEEE Access* 7 (2019), 156694–156706.

[52] Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. 2019. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1426–1435.

[53] Xingming Yang, Fei Xu, Kewei Wu, Zhao Xie, and Yongxuan Sun. 2021. Gaze-Aware Graph Convolutional Network for Social Relation Recognition. *IEEE Access* 9 (2021), 99398–99408.

[54] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *2011 International conference on computer vision*. IEEE, 1331–1338.

[55] Yu Yu, Gang Liu, and Jean-Marc Odobez. 2019. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11937–11946.

[56] Gloria Zen, Enver Sangineto, Elisa Ricci, and Nicu Sebe. 2014. Unsupervised Domain Adaptation for Personalized Facial Emotion Recognition. In *Proceedings of the 16th International Conference on Multimodal Interaction* (Istanbul, Turkey) *(ICMI '14)*. Association for Computing Machinery, New York, NY, USA, 128–135. https://doi.org/10.1145/2663204.2663247

[57] Xucong Zhang, Michael Xuelin Huang, Yusuke Sugano, and Andreas Bulling. 2018. Training person-specific gaze estimators from user interactions with

multiple devices. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[58] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. *Advances in neural information processing systems* 27 (2014).

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2242–2251. https://doi.org/10.1109/ICCV.2017.244

[60] Ning Zhuang, Bingbing Ni, Yi Xu, Xiaokang Yang, Wenjun Zhang, Zefan Li, and Wen Gao. 2019. Muggle: Multi-stream group gaze learning and estimation. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 10 (2019), 3637–3650.

[61] Andrea Zunino, Jacopo Cavazza, Riccardo Volpi, Pietro Morerio, Andrea Cavallo, Cristina Becchio, and Vittorio Murino. 2020. Predicting intentions from motion: The subject-adversarial adaptation approach. *International Journal of Computer Vision* 128, 1 (2020), 220–239.