

Detecting Attended Visual Targets in Video

Eunji Chong¹Yongxin Wang²Nataniel Ruiz³James M. Rehg¹¹Georgia Institute of Technology²Carnegie Mellon University³Boston University

{eunjichong, rehg}@gatech.edu, yongxinw@andrew.cmu.edu, nruiz9@bu.edu

<https://github.com/ejcg/t/attention-target-detection>

Figure 1: Visual attention target detection over time. We propose to solve the problem of identifying gaze targets in video. The goal of this problem is to predict the location of visually attended region (circle) in every frame, given a track of an individual’s head (bounding box). It includes the cases where such target is out-of-frame (row-col: 1-2, 1-3, 2-1), in which case the model should correctly infer its absence.

Abstract

We address the problem of detecting attention targets in video. Our goal is to identify where each person in each frame of a video is looking, and correctly handle the case where the gaze target is out-of-frame. Our novel architecture models the dynamic interaction between the scene and head features and infers time-varying attention targets. We introduce a new annotated dataset, *VideoAttentionTarget*, containing complex and dynamic patterns of real-world gaze behavior. Our experiments show that our model can effectively infer dynamic attention in videos. In addition, we apply our predicted attention maps to two social gaze behavior recognition tasks, and show that the resulting classifiers significantly outperform existing methods. We achieve state-of-the-art performance on three datasets: *GazeFollow* (static images), *VideoAttentionTarget* (videos), and *VideoCoAtt* (videos), and obtain the first results for automatically classifying clinically-relevant gaze behavior without wearable cameras or eye trackers.

1. Introduction

Gaze behavior is a critically-important aspect of human social behavior, visual navigation, and interaction with the 3D environment [26, 27]. While monitor-based and wear-

able eye trackers are widely-available, they are not sufficient to support the large-scale collection of naturalistic gaze data in contexts such as face-to-face social interactions or object manipulation in 3D environments. Wearable eye trackers are burdensome to participants and bring issues of calibration, compliance, cost, and battery life.

Recent works have demonstrated the ability to estimate the gaze target directly from images, with the potential to greatly increase the scalability of naturalistic gaze measurement. A key step in this direction was the work by Recasens *et al.* [44], which demonstrated the ability to detect the attention target of each person within a single image. This approach was extended in [11] to handle the case of out-of-frame gaze targets. Other related works include [45, 48, 28, 54, 18]. These approaches are attractive because they can leverage head pose features, as well as the saliency of potential gaze targets, in order to resolve ambiguities in gaze estimation.

This paper develops a spatiotemporal approach to gaze target prediction which models the dynamics of gaze from video data. Fig 1 illustrates our goal: For each person in each video frame we estimate where they are looking, including the correct treatment of out-of-frame gaze targets. By identifying the visually-attended region in every frame, our method produces a dense measurement of a person’s natural gaze behavior. Furthermore, this approach it has the

benefit of linking gaze estimation to the broader tasks of action recognition and dynamic visual scene understanding.

An alternative to the dynamic prediction of gaze targets is to directly classify specific categories or patterns of gaze behaviors from video [32, 40, 31, 14, 50, 15]. This approach treats gaze analysis as an action detection problem, for actions such as mutual gaze [32, 40, 31] or shared attention to an object [14, 50]. While these methods have the advantage of leveraging holistic visual cues, they are limited by the need to pre-specify and label the target behaviors. In contrast, our approach of predicting dense gaze targets provides a flexible substrate for modeling domain-specific gaze behaviors, such as the assessments of social gaze used in autism research [37, 4].

A key challenge in tackling the dynamic estimation of gaze targets in video is the lack of suitable datasets containing ground truth gaze annotations in the context of rich, real-world examples of complex time-varying gaze behaviors. We address this challenge by introducing the *VideoAttentionTarget* dataset, which contains 1,331 video sequences of annotated dynamic gaze tracks of people in diverse situations.

Our approach to spatiotemporal gaze target prediction has two parts. First, we develop a novel spatial reasoning architecture to improve the accuracy of target localization. The architecture is composed of a scene convolutional layer that is regulated by the head convolutional layer via an attention mechanism [2], such that the model focuses on the scene region that the head is oriented to. The spatial module improves the state-of-the-art result on the GazeFollow benchmark by a considerable margin. Second, we extend the model in the temporal dimension through the addition of ConvLSTM networks. This model outperforms multiple baselines on our novel VideoAttentionTarget dataset. The software, models and dataset are made freely-available for research purposes.

We further demonstrate the value of our approach by using the predicted heatmap from our model for social gaze recognition tasks. Specifically, we experimented on two tasks: 1) Automated behavioral coding of the social gaze of young children in an assessment task, and 2) Detecting shared attention in social scenes. In the first experiment, our heatmap features were found to be the most effective among multiple baselines for attention shift detection. In the second experiment, our approach achieved state-of-the-art performance on the VideoCoAtt dataset [14]. Both results validate the feasibility and effectiveness of leveraging our gaze target prediction model for gaze behavior recognition tasks. This paper makes the following contributions:

- A novel spatio-temporal deep learning architecture that learns to predict dynamic gaze targets in video
- A new *VideoAttentionTarget* dataset, containing dense

annotations of attention targets with complex patterns of gaze behavior

- Demonstration that our model’s predicted attention map can achieve state-of-the art results on two social gaze behavior recognition tasks

2. Related Work

We organize the related work into three areas: gaze target prediction, gaze behavior recognition, and applications to social gaze analysis. Our focus is gaze target prediction, but we also provide results for behavior recognition in a social gaze setting (see Secs. 5.3 and 5.4).

Gaze Target Prediction One key distinction among previous works on gaze target prediction is whether the attention target is located in a 2D image [44, 45, 48, 11, 28, 54, 18] or 3D space [1, 33, 51, 3, 34]. Our work addresses the 2D case, which we review in more detail; Authors of [44] were among the first to demonstrate how a deep model can learn to find the gaze target in the image. Saran *et al.* [48] adapt the method of [44] to a human-robot interaction task. Chong *et al.* [11] extends the approach of [44] to address out-of-frame gaze targets by simultaneously learning gaze angle and saliency. Within-frame gaze target estimation can be further enhanced by considering different scales [28], body pose [18] and sight lines [54]. A key difference between these works and our approach is that we explicitly model the gaze behavior over time and report results for gaze target prediction in video while considering out-of-frame targets.

Our problem formulation and network architecture are most closely-related to [11]. In addition to the temporal modeling, three other key differences from [11] are 1) that we do not supervise with gaze angles and not require auxiliary datasets; 2) therefore we greatly simplify the training process; and 3) we present an improved spatial architecture. In terms of architecture, 1) we use head features to regulate the spatial pooling of the scene image via an attention mechanism; 2) we use a head location map instead of one-hot position vector; and 3) we use deconvolutions instead of a grid output to produce a fine-grained heatmap. Our experiments show that these innovations result in improved performance on GazeFollow (*i.e.* for static images, see Table 1) and on our novel video attention dataset (see Table 2).

The work of [45] shares our goal of inferring gaze targets from video. In contrast to our work, they address the case where the gaze target is primarily visible at a later point in time, after the camera pans or there is a shot change. While movies commonly include such indirect gaze targets, they are rare in the social behavior analysis tasks that motivate this work (see Fig. 6). Our work is complementary to [45], in that our model infers per-frame gaze targets.

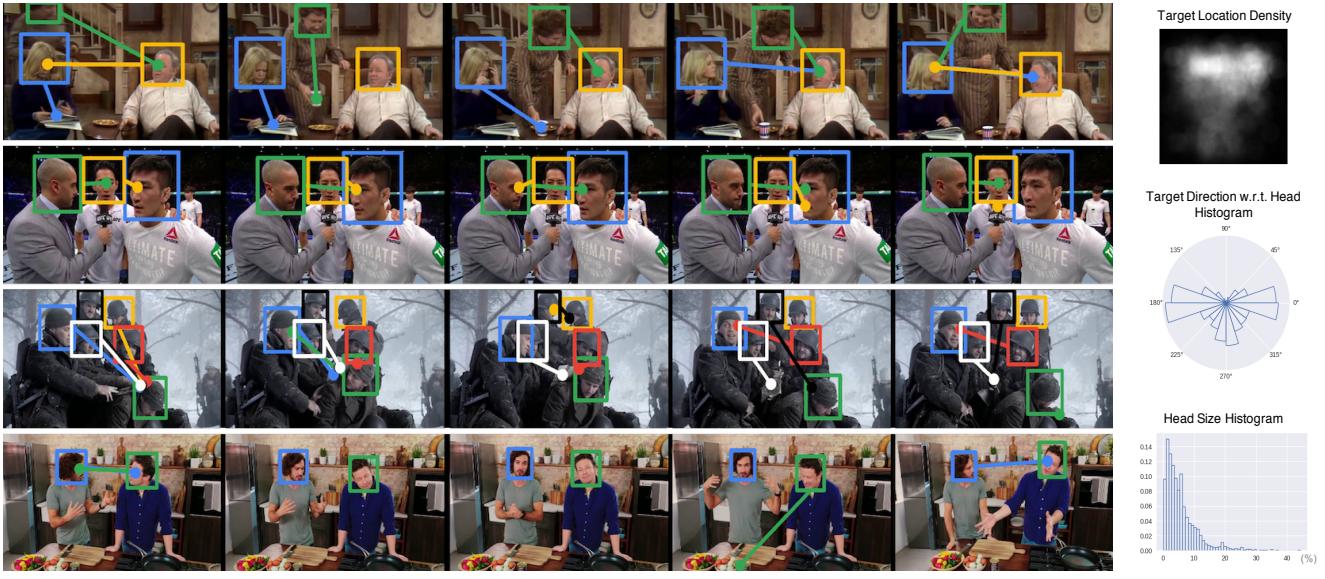


Figure 2: Overview of novel *VideoAttentionTarget* dataset (a) Example sequences illustrating the per-frame annotations of each person (bounding box) and their corresponding gaze target (solid dot). (b) Annotation statistics: top - annotated gaze target location distribution in image coordinates, middle - histogram of directions of gaze targets relative to the head center, bottom - histogram of head sizes measured as the ratio of the bounding box area to the frame size.

Several works address the inference of 3D gaze targets [1, 33, 34]. In this setting, the identification of an out-of-frame gaze target can be made by relying on certain assumptions about the scene, such as the target object’s location or its motion, or by using a joint learning framework informed by the task [51] or target location [3].

Gaze Behavior Recognition An alternative to inferring the target gaze location is to directly infer a gaze-related behavior of interest. For example, several approaches have been developed to detect if two people are looking at each other [32, 40, 31], or to detect if more than two people are looking at a common target [14, 50]. In addition, the 3D detection of socially-salient regions has been investigated using an egocentric approach [42]. Recently, Fan *et al.* [15] addressed the problem of recognizing atomic-level gaze behavior when human gaze interactions are categorized into six classes such as avert, refer, and follow.

In contrast to approaches that directly infer gaze behavior, our method provides a dense mid-level representation of attention for each person in a video. Thus our approach is complementary to these works, and we demonstrate in Secs. 5.3 and 5.4 that our gaze representation has utility for gaze behavior classification.

Social Gaze Detection in Clinical Settings One motivation for our work is the opportunity for automated measure-

ments of gaze behavior to inform research and clinical practice in understanding and treating developmental conditions such as autism [46, 19]. In this setting, automated analysis can remove the burden of laborious gaze coding that is commonplace in autism research, and enable a more fine-grained analysis of gaze behavior in clinical populations. Prior work in this area has leveraged the ability to analyze head orientation [47, 17, 12] to infer children’s attention and have developed solutions for specific settings [43, 21, 5]. Prior works have also addressed the detection of eye contact and mutual gaze in the context of dementia care [35, 39] and autism [53, 9, 10]. Other work has analyzed mutual gaze in group interactions for inferring rapport [36]. In contrast to these works, our focus is to first develop a general approach to gaze target identification in video, and then explore its utility in estimating clinically-important social behaviors during face-to-face interactions between an adult examiner and a child. We believe are the first to present results (in Sec. 5.3) for automatically detecting clinically-meaningful social gaze shifts without a wearable camera or an eye tracker.

3. VideoAttentionTarget Dataset

In this section we describe our novel *VideoAttentionTarget* dataset that was created specifically for the task of video gaze target modeling. Some example frames, annotations and statistics of the dataset are shown in Fig 2.

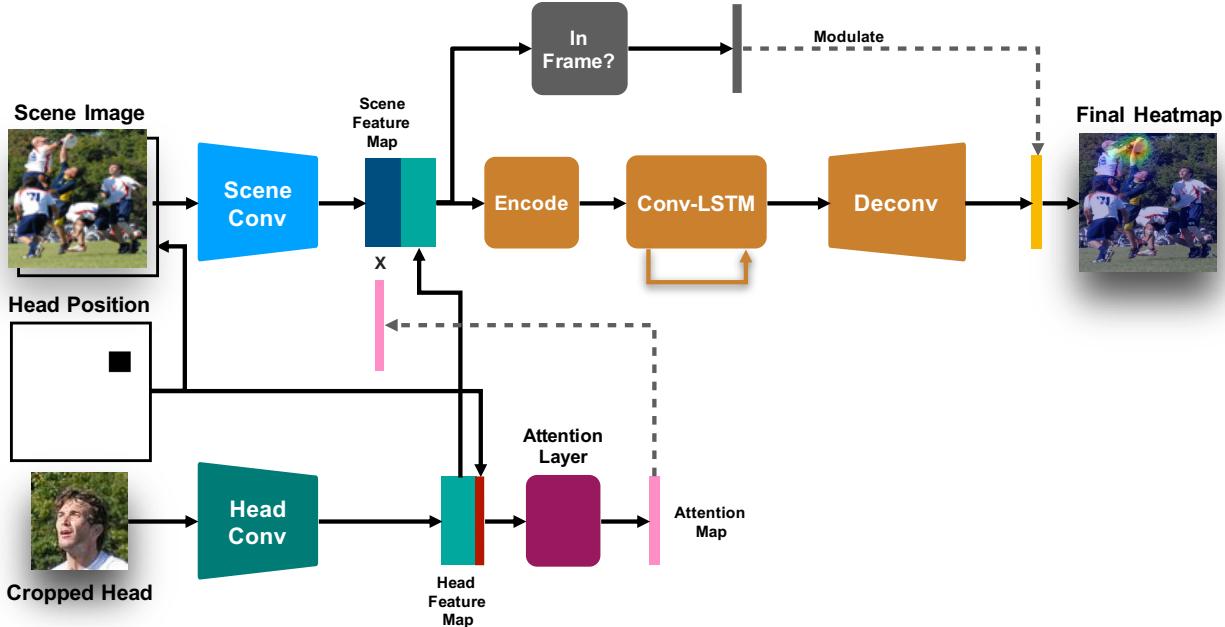


Figure 3: **Spatiotemporal architecture for gaze prediction.** It consists of a head conditioning branch which regulates the main scene branch using an attention mechanism. A recurrent module generates a heatmap that is modulated by a scalar, which quantifies whether the gaze target is in-frame. Displayed is an example of in-frame gaze from the GazeFollow dataset.

In order to ensure that our dataset reflects the natural diversity of gaze behavior, we gathered videos from various sources including live interviews, sitcoms, reality shows, and movie clips, all of which were available on YouTube. Videos from 50 different shows were selected. From each source video, we extracted short clips that contain dynamic gaze behavior without scene cuts, in which a person of interest can be continuously observed. The length of the clips varies between 1-80 seconds.

For each clip, annotators first labeled tracks of head bounding boxes for each person. This resulted in 1,331 tracks comprising 164,541 frame-level bounding boxes. In the second pass, the annotators labeled the gaze target as a point in each frame for each annotated person. They also had the option to mark if the target was located outside the video frame (including the case where the subject was looking at the camera). This produced 109,574 in-frame gaze targets and 54,967 out-of-frame gaze annotations. All frames in all clips were annotated using custom software by a team of four annotators, with each frame annotated once.

A testing set was constructed by holding out approximately 20% of the annotations (10 shows, 298 tracks, 31,978 gaze annotations), ensuring no overlap of shows between the train and test splits. This allows us to measure generalization to new scenarios and individuals. Furthermore, in order to characterize the variability in human annotations of gaze targets, we had two other annotators (among the four who annotated the train split) who did not label

that particular test samples additionally annotate them. We report this human inter-rater reliability which serves as the upper bound on the algorithm performance.

4. Spatiotemporal Gaze Architecture

Our architecture is composed of three mains parts. A **head conditioning branch**, a **main scene branch** and a **recurrent attention prediction module**. An illustration of the architecture is shown in Fig. 3.

Head Conditioning Branch The head conditioning branch computes a head feature map from the crop of the head of the person of interest in the image. The “Head Conv” part of the network is a ResNet-50 [20] followed by an additional residual layer and an average pooling layer. A binary image of the head position, with black pixels designating the head bounding box and white pixels on the rest of the image, is reduced using three successive max pooling operations and flattened. We found that the binary image encoded the location and relative depth of the head in the scene more effectively than the position encoding used in previous works. The head feature map is concatenated with this head position feature. An attention map is then computed by passing these two concatenated features through a fully-connected layer which we call the “Attention Layer”.

Main Scene Branch A scene feature map is computed using the “Scene Conv” part of the network, which is identical to the “Head Conv” module previously described. Input to the “Scene Conv” is a concatenation of scene image and

head position image. We found that providing head position as a spatial reference along with the scene helped the model learn faster. This scene feature map is then multiplied by the attention map computed by the head conditioning branch. This enables the model to learn to pay more attention to the scene features that are more likely to be attended to, based on the properties of the head. In comparison to [11], our approach results in earlier fusion of the scene and head information. The head feature map is additionally concatenated to the weighted scene feature map. Finally, the concatenated features are encoded using two convolutional layers in the “Encode” module.

Recurrent Attention Prediction Module After encoding, the model integrates temporal information from a sequence of frames using a convolutional Long Short-Term Memory network [52], designated as “Conv-LSTM” in Fig. 3. A deconvolutional network comprised of four deconvolution layers, designated as the “Deconv” module, up-samples the features computed by the convolutional LSTM into a full-sized feature map. We found that this approach yields finer details than the grid-based map used in [11].

Heatmap Modulation The full-sized feature map is then modulated by a scalar α which quantifies whether the person’s focus of attention is located inside or outside the frame, with higher values indicating in-frame attention. This α is learned by the “In Frame?” module in Fig. 3, which consists of two convolutional layers followed by a fully-connected layer. The modulation is performed by an element-wise subtraction of the $(1 - \alpha)$ from the normalized full-sized feature map, followed by clipping of the heatmap such that its minimum values are ≥ 0 . This yields the final heatmap which quantifies the location and intensity of the predicted attention target in the frame. In Fig. 3 we overlay the final heatmap on the input image for visualization.

Implementation Details We implemented our model in PyTorch. The input to the model is resized to 224×224 and normalized. The Attention Layer outputs 7×7 spatial soft-attention weights. The ConvLSTM module uses two ConvLSTM layers with kernels of size 3, whose output is up-sampled to a 64×64 -sized heatmap. Further model specifications can be found in our code.

For supervision, we place a Gaussian weight around the center of the target to create the ground truth heatmap. Heatmap loss \mathcal{L}_h is computed using MSE loss when the target is in frame per ground truth. In-frame loss \mathcal{L}_f is computed with binary cross entropy loss. Final loss \mathcal{L} used for training is a weighted sum of these two: $\mathcal{L} = w_h \cdot \mathcal{L}_h + w_f \cdot \mathcal{L}_f$.

We initialize the Scene Conv with CNN for scene recognition [55] and the Head Conv with CNN for gaze estimation [16]. Training is performed in a two-step process. First, the model is globally trained on the GazeFollow dataset until convergence. Second, it is subsequently trained on the

Method	AUC \uparrow	Average Dist. \downarrow	Min Dist. \downarrow
Random	0.504	0.484	0.391
Center	0.633	0.313	0.230
Judd [23]	0.711	0.337	0.250
GazeFollow [44]	0.878	0.190	0.113
Chong [11]	0.896	0.187	0.112
Zhao [54]	n/a	0.147	0.082
Lian [28]	0.906	0.145	0.081
Ours	0.921	0.137	0.077
Human	0.924	0.096	0.040

Table 1: **Spatial module evaluation** on the GazeFollow dataset for single image gaze target prediction.

Method	<i>spatial</i>		<i>out of frame</i>
	AUC \uparrow	L^2 Dist. \downarrow	AP \uparrow
Random	0.505	0.458	0.621
Fixed bias	0.728	0.326	0.624
Chong [11]	0.830	0.193	0.705
Chong [11]+LSTM	0.833	0.171	0.712
No head position	0.835	0.169	0.827
No head features	0.758	0.258	0.714
No attention map	0.717	0.226	0.774
No fusion	0.853	0.165	0.817
No temporal	0.854	0.147	0.848
Ours full	0.860	0.134	0.853
Human	0.921	0.051	0.925

Table 2: **Quantitative model evaluation** on our VideoAttentionTarget dataset.

VideoAttentionTarget dataset, while freezing the layers up to the Encode module to prevent overfitting. We used random flip, color jitter, and crop augmentations as described in [44]. We also added noise to head position during training to minimize the impact of head localization errors.

5. Experiments

We conducted four experiments to evaluate the performance of our method. Sec. 5.1 uses just the spatial component of our model on the GazeFollow dataset. Sec. 5.2 uses the full spatiotemporal model on the VideoAttentionTarget dataset. Sec. 5.3 uses the model output to classify clinically-relevant social behaviors in a sample of toddlers. Sec. 5.4 uses the model to detect shared attention in the VideoCoAtt dataset. *Our method produces state-of-the-art results on all datasets in all experiments.*

5.1. Spatial Module Evaluation

We evaluate the static part of our model on single image gaze target prediction using the GazeFollow dataset [44], and compare against prior methods. Evaluation follows the

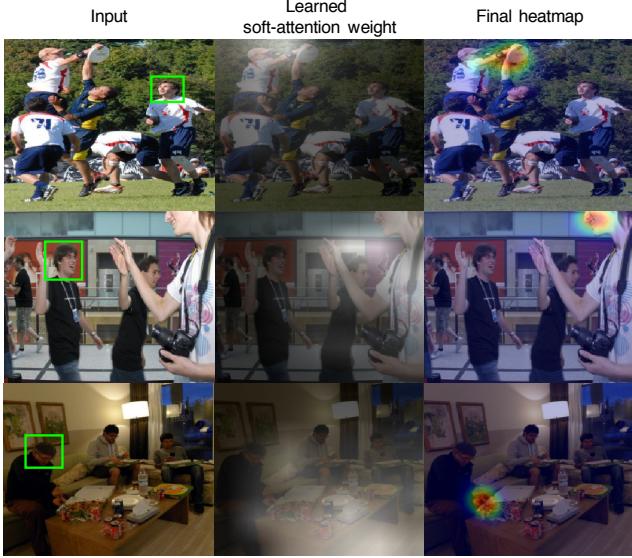


Figure 4: **Visualization of head-conditioned attention** with corresponding input and final output. The attention layer captures and leverages the head pose information to regulate the model’s prediction.

same protocol from [44, 11]. GazeFollow contains annotations of person heads and gaze locations in a diverse set of images. To train the model, we used the annotation labels from [11] which were extended to additionally specify whether the annotated gaze target is out-of-frame. In order to make a fair comparison, we only use the GazeFollow dataset for training and do not include our new dataset in this experiment.

The results in Table 1 demonstrate the value of our architectural choices in the spatial model component. We outperform previous methods by a significant margin. In fact, our AUC of 0.921 is quite close to the AUC of 0.924 obtained by human. Qualitatively, visualization of the learned weights of the attention layer reveals that the model has learned to effectively make use of the facial orientation information for weighting scene features, as shown in Fig. 4.

5.2. Spatiotemporal Model Evaluation

We evaluate our full model on the new *VideoAttention-Target* dataset. We use three performance measures: AUC, Distance, and Out-of-Frame AP. **AUC:** Each cell in the spatially-discretized image is classified as gaze target or not. The ground truth comes from thresholding a Gaussian confidence mask centered at the human annotator’s target location. The final heatmap provides the prediction confidence score which is evaluated at different thresholds in the ROC curve. The area under curve (AUC) of this ROC curve is reported. **Distance:** L^2 distance between the annotated target location and the prediction given by the pixel of maximum value in the heatmap, with image width and height normal-

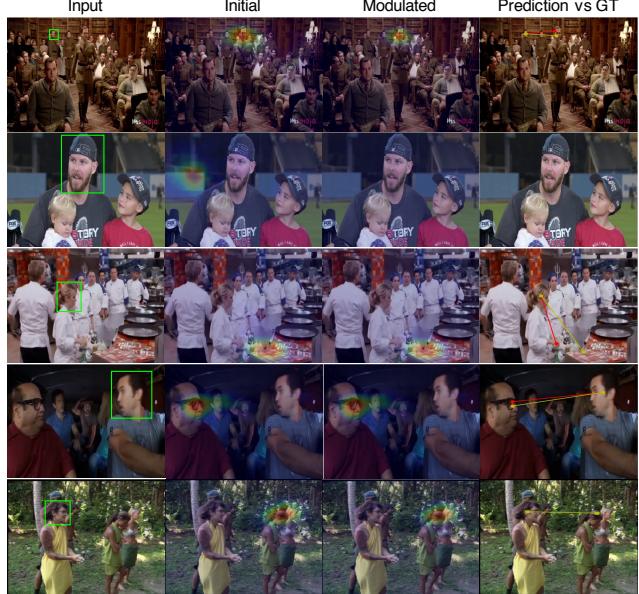


Figure 5: **Gaze target prediction results on example frames.** *Initial* denotes the first output of the deconvolution, *Modulated* shows the adjusted heatmap after modulation. Final prediction (yellow) and ground truth (red) are presented in the last column. Rows 1, 3, 4 depict properly predicted within-image gaze target, row 2 shows correctly identified nonexistent gaze target in frame, and the last row is an example of failure case where it predicts a fixated target behind the subject in the image due to the lack of sense of depth.

ized to 1. AUC and Distance are computed whenever there is an in-frame ground truth gaze target (the heatmap always has a max). **Out-of-Frame AP:** The average precision (AP) is computed for the prediction score from the scalar α (described in Sec. 4) against the ground truth, computed in every frame. We also evaluate the performance of the annotators (**Human** performance) across all three measures. This is done by comparing annotator predictions in all pairs and averaging them. This is analogous to the kappa score used to measure inter-rater reliability, but specialized for our performance measures.

Table 2 summarizes the experimental results. The first block of rows shows baseline tests and comparison with previous methods; *Random* is when the prediction is made at 50% chance, and *Fixed bias* is when the bias present in the dataset (Fig 2b) is utilized. The method of [11], which is the existing non-temporal gaze target estimator, is compared both as-is and using an additional LSTM layer on top. The second set of rows in the table shows ablation study results by disabling key components of our model one at a time; *No head position* is when the head position image is not used. *No head features* is when the head feature map

from the Head Conv module is not provided. In this case, the attention map is made using only the head position. *No attention map* is when attention map is not produced therefore the scene feature map is uniformly weighted. *No fusion* is when the head feature map is only used to produce attention map and not concatenated with scene feature map for encoding. *No temporal* is when ConvLSTM is not used. This quantitative analysis demonstrates that our proposed model strongly outperforms previous methods as well as the presented baselines. All components of the model are crucial to achieving the best performance, and the head convolutional pathway and the attention mechanism were found to have the biggest contribution. Qualitative results are presented in Fig. 5.

5.3. Detecting the Social Bids of Toddlers

Motivation Eye contact and joint attention are among the earliest social skills to emerge in human development [6], and are closely-related to language learning [22] and socio-emotional development [30]. Children with autism exhibit difficulty in modulating gaze during social interactions [38, 8, 49], and social gaze is assessed as part of the diagnosis and treatment of autism. This is usually done qualitatively or through laborious manual methods. The ability to automatically quantify children’s social gaze would offer significant benefits for clinicians and researchers.

Toddler Dataset We sampled a dataset of 20 toddlers from [9] (10 with an autism diagnosis, 10 female, mean age 36.4 months) who were video-recorded during dyadic social interactions. In this dataset, each participant completed an assessment known as the ESCS [37], which was administered by trained examiners. The ESCS is a semi-structured play protocol designed to elicit nonverbal social communication behaviors.

Five expert raters annotated all of the child’s gaze behavior consisting of looks to toys and looks to the examiner’s face at the frame level. Based on this per-frame annotation, a toy-to-eyes gaze shift event is inferred if the gaze target changes from the toy to the examiner’s face within 700 milliseconds. In total, the dataset contains 623 shift events during 221-minute-long recordings. Our task was to detect these toy-to-eyes gaze shifts and reject all other types of gaze shifts which the child made during the session. The toy-to-eyes shifts are relevant to child development because they can be further classified into different types of joint attention based on the context in which they are produced [37]. Joint attention is a key construct for the development of social communication. Our experiment provides preliminary evidence for the feasibility of automatically identifying such gaze-based joint attention events from video.

Experimental setup and results Given the toddlers



Figure 6: **Heatmap output** of our model on toddlers video.

dataset, we conducted experiments to see how an automated method can be used to retrieve gaze shift events. Two types of approaches for shift detection are explored. The first approach is to detect a shift in a two-step process where we initially classify the type of attended object - among toy, eyes, and elsewhere - in every frame with a ResNet-50 image classifier, and then apply an event classifier on top of it over a temporal window to conclusively find the gaze shift from toy to eyes. A random forest model is used for the event classifier. For the second approach, we try detecting a shift event in an end-to-end manner, using the I3D model [7] since gaze shift can be viewed as a special case of a human action.

For both approaches, we compare shift detection performance when the inputs to the models are 1. the RGB image alone, 2. image and head position, and 3. image and heatmap produced by our attention network (Fig. 6). For 2 and 3 the head position or heatmap is concatenated depth-wise to the RGB image as a 4th channel in grayscale. CNN layers of ResNet were pretrained on ImageNet [13] and those of I3D were pretrained on Kinetics [24]. The child’s head was detected and recognized using [25]. A sliding window size of 64 frames was used during training. For validation, we adopted 5-fold subject-wise cross validation in which 4 subjects were held out in each validation set.

Table 3 summarizes the results of our experiment with the precision and recall of gaze shift detection. Interestingly, the 2D-CNN-based approach generally outperformed the 3D-CNN method, which is presumably due to the complexity of I3D and relatively less training data. Nevertheless, there still exists a noticeable gap relative to human performance, implying the need for further research on this problem.

5.4. Detecting Shared Attention in Social Scenes

As an additional application of our system on real-world problems, we apply our model to infer shared attention in social scenes. We use the VideoCoAtt dataset [14] to bench-

Method	Detection Approach	Prec. \uparrow	Rec. \uparrow
Random		0.034	0.503
ResNet on RGB	random	0.541	0.567
ResNet on RGB+head	random forest	0.598	0.575
ResNet on RGB+hm		0.708	0.759
I3D on RGB		0.433	0.506
I3D on RGB+head	end-to-end	0.475	0.500
I3D on RGB+hm		0.559	0.710
Human (clinical experts)		0.903	0.922

Table 3: **Gaze coding detection results** on the toddlers dataset. As shown, our heatmap feature (denoted as **hm**) indeed improves shift detection when used along with image in a standard classification paradigm.

mark our performance on this task. This dataset has 113,810 test frames that are annotated with the target location when it is simultaneously attended by two or more people.

Given that our model does not have a head detection module as in [14], we trained a SSD-based [29] head detector in the same manner as [31] to automatically generate the input head positions. We fine-tuned this head detector with the head annotations in VideoCoAtt. However, we chose not to fine-tune our model for gaze target detection with VideoCoAtt, since their annotations do not naturally translate to the dense single-subject-target annotations that our model requires for training.

Our method is evaluated on the following two tasks: 1. location prediction (spatial) and 2. interval detection (temporal) of shared attention. For the localization task, we first add up the individual heatmaps of all people in the frame and aggregate them into a single shared attention confidence map (examples in Fig. 7). Then, the L^2 distance is computed between the pixel location of the maximum confidence and the center of the ground truth. For the interval detection task, we regard the aggregated confidence map as representing a shared attention case if its maximum score is above certain threshold. A single heatmap from our model can have a maximum value of 1 at the fixated location and when another heatmap is added to the same location its value becomes 2. We chose a threshold value of 1.8 instead of 2 in this experiment to make a room for slight misalignments between multiple fixations.

As a result, our method achieves state-of-the-art results on both tasks, as shown in Table 4. This outcome is surprising since the models of [14, 50] were formulated specifically to detect shared attention, whereas ours was not. However, it must also be noted that there exist differences in the experimental setup, such as the head detector and the training data, thus there are some caveats associated with our experimental finding. Here, we intend to demonstrate the



Figure 7: **Constructed shared attention map** obtained by adding up individual heatmaps of all people in the image. Samples are from the VideoCoAtt dataset.

Method	Accuracy \uparrow	L^2 Dist. \downarrow
Random	50.8	286
Fixed bias	52.4	122
GazeFollow [44]	58.7	102
Gaze+Saliency [41]	59.4	83
Gaze+Saliency [41]+LSTM	66.2	71
Fan [14]	71.4	62
Sumer [50]	78.1	63
Ours	83.3	57

Table 4: **Shared attention detection results** on the VideoCoAtt dataset. The interval detection task is evaluated with prediction accuracy and the localization task is measured with L^2 .

potential value of our model for recognizing higher-level social gaze, and it is encouraging that we can achieve good performance without tweaking the model for this specific problem.

6. Conclusion

We have presented a new deep architecture and a novel *VideoAttentionTarget* dataset for the task of detecting the time-varying attention targets for each person in a video. Our model is designed to allow the face to direct the learning of gaze-relevant scene regions, and our new dataset makes it possible to learn the temporal evolution of these features. The strong performance of our method on multiple benchmark datasets and a novel social gaze recognition task validates its potential as a useful tool for understanding gaze behavior in naturalistic human interactions.

7. Acknowledgement

We thank Caroline Dalluge and Pooja Parikh for the gaze target annotations in the *VideoAttentionTarget* dataset, and Stephan Lee for building the annotation tool and performing annotations. The toddler dataset used in Sec. 5.3 was collected and annotated under the direction of Agata Rozga, Rebecca Jones, Audrey Southerland, and Elysha Clark-Whitney. This study was funded in part by the Simons Foundation under grant 383667 and NIH R01 MH114999.

References

- [1] Sileye O Ba and Jean-Marc Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):16–33, 2008.
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [3] Ernesto Brau, Jinyan Guan, Tanya Jeffries, and Kobus Barnard. Multiple-gaze geometry: Inferring novel 3d locations from gazes observed in monocular video. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [4] Susan E Bryson, Lonnie Zwaigenbaum, Catherine McDermott, Vicki Rombough, and Jessica Brian. The autism observation scale for infants: scale development and reliability data. *Journal of autism and developmental disorders*, 38(4):731–738, 2008.
- [5] Kathleen Campbell, Kimberly LH Carpenter, Jordan Hashemi, Steven Espinosa, Samuel Marsan, Jana Schaich Borg, Zhuoqing Chang, Qiang Qiu, Saritha Vermeer, Elizabeth Adler, Mariano Tepper, Helen L Egger, Jeffery P Baker, Guillermo Sapiro, and Geraldine Dawson. Computer vision analysis captures atypical attention in toddlers with autism. *Autism*, pages 1–10, 2018.
- [6] Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the society for research in child development*, pages i–174, 1998.
- [7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.
- [8] Tony Charman, John Swettenham, Simon Baron-Cohen, Antony Cox, Gillian Baird, and Auriol Drew. Infants with autism: An investigation of empathy, pretend play, joint attention, and imitation. *Developmental psychology*, 33(5):781, 1997.
- [9] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M Jones, Agata Rozga, and James M Rehg. Detecting gaze towards eyes in natural social interactions and its use in child assessment. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):43, 2017.
- [10] Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L Ajodan, Melanie R Silverman, Catherine Lord, Agata Rozga, Rebecca M Jones, et al. Detection of eye contact with deep neural networks is as accurate as human experts.
- [11] 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.
- [12] Eunji Chong, Audrey Southerland, Abhijit Kundu, Rebecca M Jones, Agata Rozga, and James M Rehg. Visual 3d tracking of child-adult social interactions. In *2017 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 399–406. IEEE, 2017.
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [14] 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.
- [15] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
- [16] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. 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*, pages 255–258, 2014.
- [17] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1548–1557, 2017.
- [18] Jian Guan, Liming Yin, Jianguo Sun, Shuhan Qi, Xuan Wang, and Qing Liao. Enhanced gaze following via object detection and human pose estimation. In *26th International Conference on Multimedia Modeling*, 2019.
- [19] Jordan Hashemi, Mariano Tepper, Thiago Vallin Spina, Amy Esler, Vassilios Morellas, Nikolaos Papanikolopoulos, Helen Egger, Geraldine Dawson, and Guillermo Sapiro. Computer vision tools for low-cost and noninvasive measurement of autism-related behaviors in infants. *Autism research and treatment*, 2014, 2014.
- [20] 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.
- [21] Corey DC Heath, Hemanth Venkateswara, Troy McDaniel, and Sethuraman Panchanathan. Detecting attention in pivotal response treatment video probes. In *International Conference on Smart Multimedia*, pages 248–259. Springer, 2018.
- [22] Masako Hirotani, Manuela Stets, Tricia Striano, and Angela D Friederici. Joint attention helps infants learn new words: event-related potential evidence. *Neuroreport*, 20(6):600–605, 2009.
- [23] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009.

- [24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.
- [25] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [26] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986.
- [27] Michael Land and Benjamin Tatler. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, 2009.
- [28] 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.
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [30] Amy C MacPherson and Chris Moore. Attentional control by gaze cues in infancy. In *Gaze-Following*, pages 53–75. Psychology Press, 2017.
- [31] 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 Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019.
- [32] Manuel Jesús Marín-Jiménez, 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.
- [33] Benoît Massé, Silène Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2711–2724, 2017.
- [34] Benoit Massé, Stéphane Lathuilière, Pablo Mesejo, and Radu Horaud. Extended gaze following: Detecting objects in videos beyond the camera field of view. In *14th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2019, Lille, France, May 14-18, 2019*, 2019.
- [35] Yu Mitsuzumi, Atsushi Nakazawa, and Toyoaki Nishida. Deep eye contact detector: Robust eye contact bid detection using convolutional neural network. In *BMVC*, 2017.
- [36] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, page 31. ACM, 2018.
- [37] Peter Mundy, Christine Delgado, Jessica Block, Meg Venezia, Anne Hogan, and Jeffrey Seibert. Early social communication scales (escs). *Coral Gables, FL: University of Miami*, 2003.
- [38] Peter Mundy, Marian Sigman, Judy Ungerer, and Tracy Sherman. Defining the social deficits of autism: The contribution of non-verbal communication measures. *Journal of child psychology and psychiatry*, 27(5):657–669, 1986.
- [39] Atsushi Nakazawa, Yu Mitsuzumi, Yuki Watanabe, Ryo Kurozume, Sakiko Yoshikawa, and Miwako Honda. First-person video analysis for evaluating skill level in the humanitude tender-care technique. *Journal of Intelligent & Robotic Systems*, pages 1–16, 2019.
- [40] Cristina Palmero, Elsbeth A van Dam, Sergio Escalera, Mike Kelia, Guido F Lichtert, Lucas PJJ Noldus, Andrew J Spink, and Astrid van Wieringen. Automatic mutual gaze detection in face-to-face dyadic interaction videos. *Measuring Behavior 2018*, 2018.
- [41] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O’Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016.
- [42] Hyun Soo Park, Eakta Jain, and Yaser Sheikh. 3D social saliency from head-mounted cameras. In *Advances in Neural Information Processing Systems*, volume 1, pages 422–430, 2012.
- [43] Guido Pusiol, Laura Soriano, Li Fei-Fei, and Michael C Frank. Discovering the signatures of joint attention in child-caregiver interaction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 36, 2014.
- [44] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015.
- [45] 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.
- [46] James M. Rehg, Agata Rozga, Gregory D. Abowd, and Matthew S. Goodwin. Behavioral Imaging and Autism. *IEEE Pervasive Computing*, 13(2):84–87, 2014.
- [47] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018.
- [48] Akanksha Saran, Srinjoy Majumdar, Elaine Schaertl Shor, Andrea Thomaz, and Scott Niekum. Human gaze following for human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8615–8621. IEEE, 2018.
- [49] Atsushi Senju and Mark H Johnson. Atypical eye contact in autism: models, mechanisms and development. *Neuroscience & Biobehavioral Reviews*, 33(8):1204–1214, 2009.
- [50] Omer Sumer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Attention flow: End-to-end joint attention estimation. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020.
- [51] Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6801–6809, 2018.
- [52] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation

- nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [53] Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M Rehg. Detecting bids for eye contact using a wearable camera. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.
- [54] Hao Zhao, Ming Lu, Anbang Yao, Yurong Chen, and Li Zhang. Learning to draw sight lines. *International Journal of Computer Vision*, pages 1–25, 2019.
- [55] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.