Connecting Gaze, Scene and Attention

Anonymous ECCV submission

Paper ID 1002

Abstract. This paper addresses the new challenging problem of estimating general visual attention of people in images. Our proposed method is designed to work across multiple naturalistic social scenarios and provides a full picture of the subject's attention and gaze, unlike most existing work on gaze and attention estimation which has focused on constrained versions of this problem in predetermined contexts. To solve this problem, we leverage three different datasets using a multi-task learning approach to learn problems that are closely related with each other regarding human visual attention. We evaluate our method on widely used benchmarks for single-tasks such as gaze angle estimation and attention-within-an-image as well as on the new challenging task of generalized visual attention prediction. Additionally we create annotations for the MMDB and GazeFollow datasets which are used for training and evaluating this new task and publicly release our added annotations.

Keywords: Visual Attention, Gaze Estimation, Saliency, Fixation

1 Introduction

As humans, we are exquisitely sensitive to the gaze of others. We can rapidly infer if another person is making eye contact, follow their gaze to identify their gaze target, categorize quick glances to objects, and even identify when someone is not paying attention [1]. Automatically detecting and quantifying these types of visual attention from images and video remains a complex, open challenge. Although gaze estimation has long been an active area of research, most work has focused on relatively constrained versions of the problem in specific predetermined contexts. For example, [2, 3] predict the gaze target *given* that the person is looking at a point on a smartphone screen, [4] predicts fixation on an object *given* that the person is looking at salient object within the frame, [5, 6] predict eye contact *given* that the camera is located near the subject's eyes, and [7] predicts the focus of a person's gaze across views in carefully filmed commercial movies which include camera views that follow the actor's attention. It remains a significant challenge to design a system that can model the visual attention of subjects in unconstrained scenarios, without the preconditions utilized by prior works. We call this the problem of **generalized visual attention prediction**.

The three examples in Figure 1 illustrate the difficulty of the challenge. In Figure 1**(a)**, the subjects are looking at a salient object in the scene, while in **(b)** the subject is looking somewhere outside of the scene, and **(c)** they are looking at the camera. The case of Figure 1**(a)** is addressed by the pioneering

Fig. 1. We present a model which aims to understand different aspects of the generalized attention prediction problem which are exemplified above. In **(a)** subjects are looking at a salient object in the scene, in **(b)** the subject is looking somewhere outside of the scene and in **(c)** the subject is looking at or around the camera. Our model sets out to predict the 3D gaze vector of subjects in each of these images as well as the location of the gaze fixation in the image, if it exists. If not, the subject is looking somehwere outside of the scene and the model should predict this case as well.

work of Recasens et al. [4], which tackles this problem by obtaining human annotations of subject gaze targets, leveraging the finding from [8–10] which indicate that annotators very often agree on which object is salient in the scene. Their approach, however, was not designed to handle cases (b) and (c) since the dataset annotation process forces human annotators to label a point in the image as the fixation location. In other words the dataset does not distinguish between subjects looking at a point inside of the image or looking somewhere outside of the image. A purely saliency based approach would also fail: notice that there are salient objects in (b), an American flag, and (c), a mug, which can confound such an approach.

Figure 1-(c) resembles the case of Krafka et al. [3] and has been tackled in other works as well thanks to the creation of datasets in which subjects look at objects in screens and where all elements are calibrated such that the direction of the gaze can be obtained in a reliable manner [11]. Figure 1-(b) presents the hardest case - it is very difficult to annotate similar images since 3D gaze direction is required. For example one way to do this is to map the 3D scene as well as capture the 3D location of the observed point in order to obtain the gaze direction. Naturally, we have not found a dataset which addresses this aspect of the generalized attention prediction problem.

Indeed, obtaining a system that can deal with these different cases is a hard problem. This paper sets out to solve this by presenting a novel generalized visual attention estimation method which jointly learns a subject-dependent saliency map and the 3D gaze vector represented by yaw and pitch. This allows us to estimate the final fixation likelihood map.

We design the method to produce the output in such a way that the fixation likelihood map becomes close to zero when the subject is looking outside the frame as in cases (b) and (c). When the subject is looking at a point inside of

the scene, similar to case (a), then the fixation likelihood map predicts where the subject is likely to attend in the image. The model simultaneously estimates the 3D gaze angle to provide a complete picture of the subject's attention and gaze. As a result, our approach produces interpretable results spanning all three cases in Figure 1.

Our Contribution. The major contribution of this work is the attention estimation method which addresses the **generalized visual attention prediction problem** and which works accross most natural scenarios. To effectively train our model we exploit three public datasets that have been originally collected for different tasks. Specifically, we use the EYEDIAP dataset [11] to learn precise gaze angle representation, a modified version of GazeFollow dataset [4] to learn gaze-relevant scene saliency representation, and the SynHead dataset [12] to complement the first two datasets as it includes large face pose variations and subject attention outside of the image.

As a result of our multi-task learning approach, our model achieves state-of-the-art results on the GazeFollow [4] task, which consists in identifying the location of the scene the subject is looking at. Our model also competes with state-of-the-art models on the 3D gaze estimation task on the EYEDIAP dataset [11]. Most importantly, we evaluate our full model on a new challenging task that automatically quantifies dense visual attention in naturalistic social interactions. We report our results on the Multimodal Dyadic Behavior (MMDB) dataset [13], a dataset of video recordings of the social and communicative behavior of toddlers. This dataset has frame-level annotations of subject's visual targets among many other nonverbal behaviors. We are the first to report attention estimation results on this dataset. We compare our results to a variety of baseline tests that are outperformed by our method.

2 Related work

Gaze Estimation: Gaze estimation aims to predict the gaze of a human subject. Our work is related to third person gaze estimation and tracking methods which seek to estimate either the three dimensional direction of the gaze or the fixation point of the gaze on a screen. Krafka et al. [3] predict the coordinates of the gaze of a person on a smartphone device and present a dataset which addresses this problem. Mora et al. [11] present EYEDIAP, a dataset designed for the evaluation of gaze estimation task and collected in a controlled lab environment. They implement a RGB-D method which predicts the 3D vector of the gaze of the subject. There exist datasets which address the same task such as MPIIGaze [2] as well as synthetic datasets of eye images for gaze estimation [14, 15]. In addition to predicting the 3D gaze vector our work predicts a fixation likelihood map of the scene as well as whether the person is looking at a location inside or outside of the image.

Visual Saliency: The objective of visual saliency prediction is to estimate locations in an image which attract the attention of humans looking *at* the image. Since the seminal work of Itti et atl. [16] visual saliency prediction has been

extensively studied. Recently deep learning methods have shown superior performance on this task due to their ability to learn features and to incorporate both local and global context into the prediction [17–19]. Our work in generalized visual attention prediction is influenced by the task of visual saliency since people tend to look at salient objects inside a scene, yet it is distinct because we consider cases where the subject is not looking at any object in the scene. A method driven mainly by saliency detection would not be able to succeed in these scenarios. Furthermore, a person actively involved in a scenario will often be looking at a different location than a human who is free-viewing an image.

Gaze Following: The paper by Recasens et al. [4] presents a new computer vision problem which deeply inspired our work. The problem can be described as follows: given a single image containing one or more people, predict the location that each person in the scene is looking at. Their work presents a novel dataset which contains human annotations of where the subjects are looking at inside of the scene. Our work differs from this previous work in that we consider cases where the subjects are looking outside of the scene and we predict these cases. In addition to predicting a fixation likelihood map for the image we predict the 3D gaze vector of the subject. Gorji and Clark [20] study a problem in the intersection of visual saliency and gaze following which consist in incorporating signals from other regions of the image which push attention to a certain part of the image. For example when subjects in an image look at a salient object in the image this amplifies the apparent saliency of the object. Again, our problem differs in that we do not predict visual saliency but we predict gaze fixation and gaze direction of subjects inside of the scene.

Attention Modeling: People have proposed different methods for measuring third-person visual attention from environment-mounted camera. By assuming body or head orientation is a good proxy to approximate visual orientation, [21] projects attention on the street by tracking pedestrians in 3D, [22, 23] model the focus of attention in crowded social scene, and [24] predicts the object in the scene that a person is interacting with which is usually indicated by hand manipulation or pointing. Our work is certainly related, although it differs because we explicitly consider the gaze of the subject.

3 Method

Figure 2 is an overview of our deep neural network model and its input and output. The model takes three inputs: the whole image, a crop of the subject's face, and the location of her face. Given the input, the model estimates 1. the subject's gaze angle in terms of yaw and pitch degrees ("where" component of visual attention), 2. the subject-dependent saliency in terms of a heatmap ("what" component of visual attention), and 3. how likely the subject is fixating at the estimated gaze target in the scene (overall "strength" of visual attention).

The model has two fully-convolutional pathways, one connected to the whole image (Figure 2-a) and the other connected to the face image (Figure 2-b). The reasoning behind having two separate pathways is inspired by the way humans

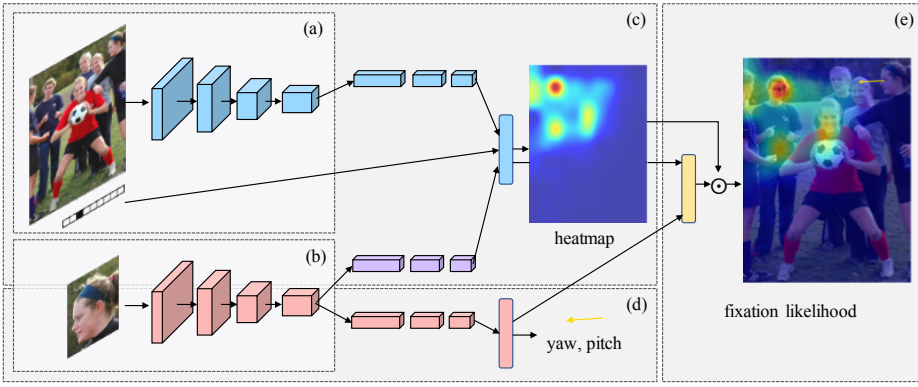

Fig. 2. Overview of our approach. Full scene image, a person's face location whose visual attention we want to predict, and the corresponding close-up face image is provided as input. Scene and face images go through separate convolutional layers in such a way that (a) (b) and (c) contribute to person-centric saliency, and (b) and (d) contribute to gaze angle prediction. In the very last layer, the final feature vectors for these two tasks are combined to estimate how likely the person is actually fixating at a gaze target within the observable scene.

infer one's visual attention, as first suggested by [4]. For example, when we interpret a person's attention from an image, we generally guess where the person's gaze is directed toward and consider if there is any salient part in the scene that lies along the direction. Based on this hypothesis, [4] connect two independent conv pathway together to learn the heatmap (Figure 2-c). We take this idea further and extend their work by explicitly training for gaze angle (Figure 2-d) with the convolutional pathway that is connected to the face image, in a multi-task learning framework. Adding the gaze angle output as an auxiliary task has several advantages including the additional supervisory signal that we can devise based on the relationship between gaze heatmap and angle, which pushes performance in heatmap estimation even further.

Lastly, we define the likelihood of fixation: a single-valued measure of how likely the subject is looking at the estimated target region inside the frame. It is modeled by a fully connected layer (Figure 2-e). Using this last output, the model can now produce a much more complete estimation of a person's visual attention. Think of the case of Figure 1-b or c, where the person is looking outside the image frame. In such cases, we want the heatmap to be as close to zero as possible since the person is not attending to any point inside of the image. By training this last layer to produce higher value for when it is more certain that the heatmap region is attended to and lower value otherwise, the value can be applied to the heatmap with an operator $\odot$ which can be a weighting operator or a gating operator depending on application.

Since there exists no single dataset that covers all various gaze and scene compositions that we consider in this paper (e.g., looking outside the frame,

looking around the camera, clear fixation on an object, etc), we adopt a cross domain learning approach where the model learns partial information relevant to each task from different datasets. Depending on what supervisory signal is available in a given batch of training data, the model selectively updates its corresponding branches.

We describe the model architecture in more detail in subsection 3.1. We elaborate on the loss function in 3.2 and talk about the datasets and training procedures in 3.3.

3.1 Model

The inputs given to the model are the entire image, the subject's cropped face and the location of the subject's face whose attention we want to estimate. The two images are resized to 227×227 so that the face can be observed in higher resolution by the network. Face position is available in terms of the (x, y) full image coordinates. These coordinates are quantized into a 13×13 grid and then flattened to a 169 dimensional 1-hot vector.

The model consists of two convolutional pathways: a face pathway (Figure 2-d) and a scene pathway (Figure 2-c). ResNet 50 [25] is used as a backbone network for the convolutional pathways (Figure 2-a,b). Specifically we use all convolutional layers of ResNet50 for each of our convolutional pathways. After each ResNet50 block we add three conv layers (1x1, followed by 3x3, followed by 1x1) with ReLu and Batchnorm - with stride 1 and no padding. The blue conv layers represented in (c) have filter depth of 512, 128 and 1 respectively. The purple and red conv layers after the face pathway (represented in (c) and (d)) have filter depth of 512, 128 and 16. These conv layers serve to reduce the dimensionality of the features extracted by the ResNet50 backbone networks.

In the face pathway, the feature vector computed with the face input image goes through a fully connected layer to predict the gaze angle represented using yaw and pitch intrinsic Euler angles. In the scene pathway, the feature vectors extracted from the whole image as well as from the face image are concatenated with the face position input vector to learn the person-centric heatmap. Similarly to face position, the ground truth used for learning the heatmap is available as a gaze target position in (x, y) coordinates which is quantized into a 10×10 grid and then flattened to a 100-dimensional one-hot vector for training.

Lastly, the input vectors to the last layer of each pathway are concatenated and go into the final fully connected layer to estimate the "strength" of the fixation ie. how likely it is that the person is actually fixating at a gaze target within the observable scene. The training label for this value is equal to 1 for a fixation inside of the image and 0 when the subject is looking outside of the scene. We also explore alternative model architectures and restrict our training to a subset of the three datasets. Experiments are reported in Section 4.4.

3.2 Loss

As our model predicts gaze angle, saliency map and the fixation likelihood, we need to apply appropriate loss functions for each task. For the angle regression

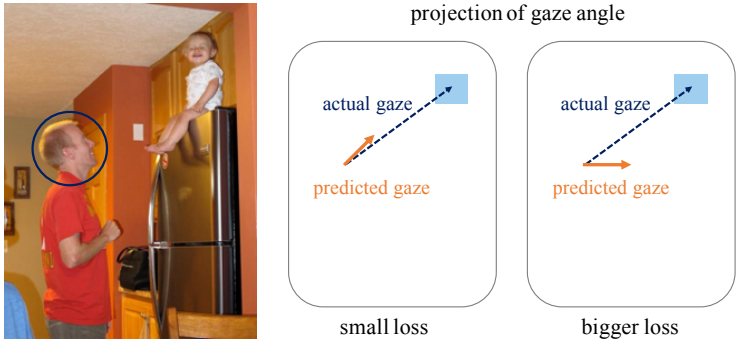

Fig. 3. Our project and compare loss is illustrated here. If the estimated angle is close to the actual one, the projected gaze angle on the image should also be close to the vector connecting the head position to the gaze target.

task we use an **L1 loss**, and for the other two tasks we use a **cross entropy loss**. Moreover, we recognize that the gaze angle and fixation target predictions are closely related. Based on their relationship additional constraints can be imposed to augment the training loss signal. Namely, when the subject is looking at a target, the actual gaze is a ray coming from the subject's head to the gaze target. This ray can be projected onto the image. It becomes a 2D vector coming from the subject's head to the target exemplified by the blue vector in Figure 3. If the estimated angle is close to the actual one, the projected gaze angle on the image (orange vector in Figure 3) should also be close to the blue vector. The closeness between the two vectors is measured using a cosine distance which supplements our learning procedure. We call this the **project and compare loss**.

3.3 Cross-Domain Datasets and Training Procedure

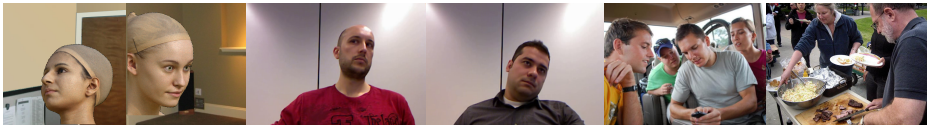

Fig. 4. Examples of datasets used to train our model. Left two: SynHead, middle two: EYEDIAP, right two: GazeFollow.

The largest challenge in training our model is the availability of training examples. Although there are a couple of existing datasets that are suitable for training certain parts of our network, no unique dataset contains all of the information that we need to train the full model. Therefore, we leverage three different

types of datasets, namely, GazeFollow [4], EYEDIAP [11], and SynHead [12]. We selectively train different sub-parts of our network at a time depending on the available supervisory information within a training batch. See Figure 4 to see sample images from each dataset.

GazeFollow [4] is a real-world dataset with manual annotations of the location of where people in images are looking. The images used in this dataset come from other major datasets such as MS COCO [26] and PASCAL [27]. As a result, the images cover a wide range of scenes, people, and gaze directions. However, the actual 3D gaze angles are not available. Furthermore, images where subjects are looking outside of the image frame are not distinguished and all images have a fixation annotation inside of the frame. Although it is mentioned in [4] that if the annotators indicated that the person was looking outside the image, the image would be discarded, we notice that there are a considerable number of images in which persons are looking outside of the frame. Therefore, we added additional annotations to this dataset in the form of a binary indicator label for "looking inside" or "looking outside" for every image. In total, we identified 14,564 images correspond to the "looking outside" case which is approximately 11.6% of the total training samples. We publicly release our additional annotations along with this paper.

EYEDIAP [11] dataset is designed for the evaluation of the gaze estimation task. It has videos of 16 different subjects with full face and background visible in a laboratory environment. Each subject was asked to look at a specific target point on a monitor screen and the 3D gaze angle was annotated by leveraging camera calibration and face depth measurement from depth camera. This dataset contains precise 3D gaze angles for frames where the person is fixating the target point. The dataset also contains video of the subjects looking at a 3D ball target instead of 2D screen target point, but we exclude these ball sessions from our experiments in order to conduct a fairer comparison with prior work. We randomly hold out four subjects for test and use the rest of the sessions for training. Since subjects were looking at a screen, all of the frames can be considered as looking outside the image. However, since the dataset has been collected in a controlled settings the background is mostly white and clear and there is not a lot of variety in lighting or pose. Also, measured gaze angles range between $-40°$ to $40°$ which is rather limited.

NVIDIA SynHead [12] is a synthetic dataset created for the head pose estimation task. The dataset contains 510,960 frames of 70 head motion tracks rendered using 10 individual head models. The gaze of the head is fixed and aligned with the head pose, thus we use the labeled 3D head pose as the gaze angle ground truth. One of the advantages of a synthetic dataset is the ability to insert different images in the background. We randomly generated 15% from the total frames augmented with provided natural scene backgrounds and regard all as "looking outside" examples. The main reason we include SynHead in training is because it complements the EYEDIAP dataset, as the angle ranges are larger, between $-90°$ and $90°$, and it can include more diverse backgrounds. Since head pose estimation is not a focus of this paper we do not set aside a test set and use SynHead entirely for training. Dataset details are also summarized in Table 1.

Table 1. Datasets used in our experiments and the number of samples in the training and testing split, as well as the percentage of each split containing peopole looking in/out.

Dataset	Training set		Test set	
		in vs out		in vs out
GazeFollow [4]	125,557	88.4% vs 11.6%	4,782	100% vs 0%
EYEDIAP [11]	72,613	0% vs 100%	18,153	0% vs 100%
SynHead [12]	75,400	0% vs 100%	-	-
MMDB [13]	-	-	4,965	41.4% vs 58.6%

Training Procedure. Since each dataset is relevant only to certain subtasks, we only update the relevant parts of the network based on which dataset the training sample is from, while freezing other irrelevant layers during backpropagation. Specifically, when learning gaze angle estimation, we only update the angle pathway (b) and (d) in Figure 2, when learning saliency we update the scene pathway (a), (b) and (c) while freezing all other layers. Similarly, when training fixation likelihood we only update the layer (e) in Figure 2. We found that this selective backpropagation scheme is critical in achieving good performance.

In every batch, we draw random samples from all of the datasets shuffled together and performe three separate backpropgataions for the three outputs as just described. In the beginning, both convolutional pathways were initialized using a ResNet50 model pretrained on the ImageNet classification task [28]. We use the Adam optimization algorithm with a learning rate of $2.5e-4$ and a batch size of 36. Training usually converges within 12 epochs. All of our implementation and experiments are done in PyTorch [29].

4 Evaluation

In this section we evaluate our results by comparing each output with a number of existing methods and baselines. We first evaluate the person-dependent saliency map in 4.1, gaze angle estimation in 4.2 and general attention estimation in 4.3. Lastly, we evaluate our method by changing model architectures and training dataset in 4.4.

4.1 Person-Dependent Saliency Prediction

We evaluate the performance of saliency map estimation using the suggested test split of the GazeFollow dataset. The test split contains all "looking inside" cases and each test image has multiple gaze target annotations. Following the same evaluation method by [4], we compute the Area Under Curve (AUC) score of the Receiver Operating Characteristic (ROC) curve in which the ground truth

Fig. 5. Qualitative results of our model's gaze-saliency prediction on the GazeFollow dataset. Input image is given on the 1st and 3rd row, the output heatmap and estimated gaze is overlaid below.

Table 2. Gaze-saliency evaluation on the GazeFollow test set

Method	AUC	L2 Distance	Min Distance	Angle
Random	0.504	0.484	0.391	69°
Center	0.633	0.313	0.230	49°
Judd [30]	0.711	0.337	0.250	54°
GazeFollow [4]	0.878	0.190	0.113	24°
Our	**0.896**	**0.187**	**0.112**	**23°**

target positions are the true labels and heatmap value on corresponding positions are prediction confidence score. Our method achieves a score of 0.896 achieving state-of-the-art performance. Along with AUC we also report results in L2, min distance and angle metric. Please refer to [4] for details about the metric. The numbers are summarized in Table 2 along with a number of baselines reported in [4]. Qualitative results are presented in Figure 5.

4.2 Gaze Angle Prediction

We report the 3D gaze estimation accuracy based on the yaw and pitch output of our model on the chosen EYEDIAP test split. Table 3 shows the angular errors in which we achieve less than 0.5 degrees of difference to the state-of-the-art appearance-based gaze estimation method. It is worth noticing that the middle two values come from [31] which are computed by five-fold cross validation with the entire EYEDIAP dataset whereas our method is evaluated on a single train/test split. Although we did not choose to perform full cross validation, we

Table 3. Gaze angle evaluation on EYEDIAP

Method	Angular Error (degree)
Wood [14]	11.3°
iTracker [3]	8.3°
Zhang [31]	**6.0°**
Our	6.4°

conclude that it reaches reasonable accuracy on the benchmark. Note also that our method is trained on multiple tasks whereas all other methods are trained solely on the gaze angle prediction task.

4.3 Generalized Attention Prediction During Naturalistic Social Interactions

The primary inspiration for our work stems from the need for the ability to quantify various types of visual attention behavior, which is one of the most important nonverbal social cues used in our daily life. Moreover, this is of particular interest among researchers who study child development since gaze behavior of young children is closely related to their social development and developmental disorders such as Autism [32]. The MMDB dataset is one of the largest datasets that contains children's social and communicative behaviors, collected in order to facilitate data-driven analysis of child behavior based on video. The dataset contains a wide range of nonverbal behavior such as hand gestures, smile, and gaze. It has frame-level human annotations of each behavior. As for gaze, each frame is annotated when the child is looking at a ball, book or the examiner. This is done by human annotation based on multiple views, therefore the child's gaze target can be visible or not depending on the viewpoint. Since the annotation does not indicate in which view the gaze target is visible, we added additional annotation ourselves and identified if the target is visible in a child-facing camera view to construct labels for the general attention estimation problem. We publicly release this annotation text file along with our paper.

We evaluate our method on the generalized attention prediction task. We design a gaze target grid classification task, where each test image is divided into N×N grids. If the subject is looking inside of the image then the grid square which contains the gaze target is assigned a label of 1 while others are assigned labels of 0. If the subject is looking somewhere outside of the frame then all grid squares are assigned the 0 label. Using our method's fixation likelihood map we predict the positive gaze grid square. We test the GazeFollow model [4] which is the closest work to our method in terms of having the ability to predict gaze target location. One of its limitations is the inability to correctly predict the "outside" case, where the subject is looking outside of the frame. As a result, our method achieves much higher precision in addition to increased recall as shown in Table 4.

Additionally, we constructed various baseline tests consisting of a classifier based on a subset of features constructed for saliency, gaze and head pose. Specifically, we tested with SVM and Random Forest using a subset of {[4], [31], [33]} as features. In other words, each classifier has been trained for detection of looking inside with the training set described in Table 1, using one or more of the three methods' output, and tested on the MMDB images. We report the results in Table 5. Note that the MMDB dataset was not used for training across all methods including ours.

Table 4. Evaluation on MMDB - gaze target grid classification

Grid Size	Method	Precision	Recall
2x2	GazeFollow [4]	0.344	0.715
	Our	0.744	0.851
5x5	GazeFollow [4]	0.210	0.437
	Our	0.614	0.683

Table 5. Evaluation of fixation likelihood on MMDB

Method	Average Precision
SVM with GazeFollow [4]	0.311
SVM with GazeFollow [4]+gaze [31]	0.531
SVM with GazeFollow [4]+headpose [33]	0.620
SVM with gaze [31]+headpose [33]	0.405
SVM with GazeFollow [4]+gaze [31]+headpose [33]	0.624
Random Forest with GazeFollow [4]	0.707
Random Forest with GazeFollow [4]+gaze [31]	0.727
Random Forest with GazeFollow [4]+headpose [33]	0.785
Random Forest with gaze [31]+headpose [33]	0.512
Random Forest with GazeFollow [4]+gaze [31]+headpose [33]	0.773
Our	**0.902**

4.4 Alternative Model and Diagnostics

Finally, we run additional experiments to study how the performance of our model is affected by different training datasets and architectural choice. As shown in Table 6, omitting EYEDIAP or SynHead training dataset did not have much impact on the heatmap estimation whereas changing model architecture considerably affected the scores. For example using a single ResNet50

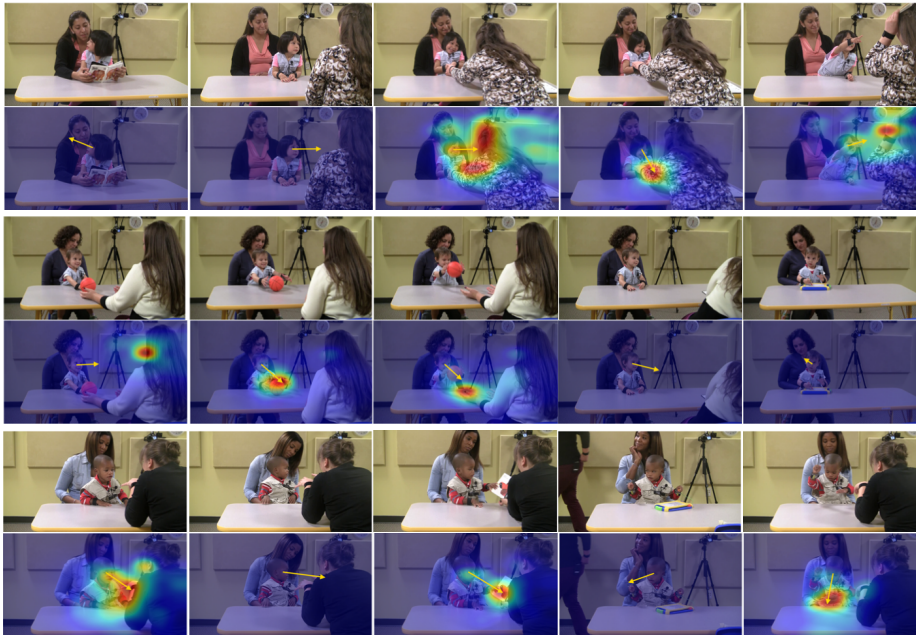

Fig. 6. Example result of our method on the MMDB dataset. The dataset contains various types of gaze behavior including fixations on a target both within and out of frame. Our method produces low heatmap when the fixation target is outside and high heatmap when the target becomes clear.

pathway which pools facial features using ROI-pooling shows significantly degraded performance which supports our decision to use a scene pathway as well as a face pathway. Qualitatively, we were able to observe that, even though our method is designed to measure fixation outside, it can make mistakes when the target is within the frame but occluded by other object. Also, when the subject is closer to the camera than some salient object in the background, the method sometimes estimates those as fixation candidate due to the lack of scene depth understanding. Examples are illustrated in Figure 7.

5 Conclusion

In this paper we present the new challenging problem of generalized visual attention prediction which encapsulates several constrained attention prediction and gaze estimation problems that have been the focus of many previous works. We propose a multi-task learning approach and neural architecture leveraging three different datasets which tackles this problem and works across multiple naturalistic social scenarios. In order to train our architecture we have supplemented these datasets with new annotations and we release these annotations to the public. Our model achieves state-of-the-art performance on the single-task

Table 6. Additional model evaluation and diagnostics

Method	GazeFollow test AUC	L2 Distance
No EYEDIAP	0.887	0.197
No SynHead	0.895	0.191
No EYEDIAP and SynHead	0.891	0.194
No project-and-compare loss	0.895	0.189
Map resolution 15x15	0.778	0.194
ROI-pooling	0.700	0.325
Our final	**0.896**	**0.187**

Fig. 7. Challenging cases due to occlusion and the lack of depth understanding.

gaze-saliency prediction and competes with state-of-the-art methods on gaze estimation benchmarks while achieving promising performance on the generalized attention prediction problem on the MMDB dataset. Future work in this area can lead to breakthroughs in attention prediction applications which are valuable in numerous scientific and commercial areas. A suggested first step would be to improve existing datasets with additional annotations or collect datasets tailored for this problem.

References

1. Land, M., Tatler, B.: Looking and acting: vision and eye movements in natural behaviour. Oxford University Press (2009)
2. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2015) 4511–4520
3. Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., Torralba, A.: Eye tracking for everyone. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
4. Recasens*, A., Khosla*, A., Vondrick, C., Torralba, A.: Where are they looking? In: Advances in Neural Information Processing Systems (NIPS). (2015) * indicates equal contribution.
5. Chong, E., Chanda, K., Ye, Z., Southerland, A., Ruiz, N., Jones, R.M., Rozga, A., Rehg, J.M.: 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) (2017) 43
6. Zhang, X., Sugano, Y., Bulling, A.: Everyday eye contact detection using unsupervised gaze target discovery. In: 30th Annual Symposium on User Interface Software and Technology, ACM (2017)
7. Recasens, A., Vondrick, C., Khosla, A., Torralba, A.: Following gaze in video. In: The IEEE International Conference on Computer Vision (ICCV). (Oct 2017)
8. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 280–287
9. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Transactions on Image Processing **24**(12) (2015) 5706–5722
10. Borji, A., Sihite, D.N., Itti, L.: What stands out in a scene? a study of human explicit saliency judgment. Vision research **91** (2013) 62–77
11. Funes Mora, K.A., Monay, F., Odobez, J.M.: Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In: Proceedings of the ACM Symposium on Eye Tracking Research and Applications, ACM (March 2014)
12. Gu, J., Yang, X., De Mello, S., Kautz, J.: Dynamic facial analysis: From bayesian filtering to recurrent neural network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (July 2017)
13. Rehg, J., Abowd, G., Rozga, A., Romero, M., Clements, M., Sclaroff, S., Essa, I., Ousley, O., Li, Y., Kim, C., et al.: Decoding children's social behavior. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2013) 3414–3421
14. Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 3756–3764
15. Sugano, Y., Matsushita, Y., Sato, Y.: Learning-by-synthesis for appearance-based 3d gaze estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1821–1828
16. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on pattern analysis and machine intelligence **20**(11) (1998) 1254–1259

17. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE (2015) 3183–3192
18. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Conference on Computer Vision and Pattern Recognition. (2015)
19. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1265–1274
20. Gorji, S., Clark, J.J.: Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 2510–2519
21. Benfold, B., Reid, I.: Guiding visual surveillance by tracking human attention. In: British Machine Vision Conference. (September 2009)
22. Soo Park, H., Shi, J.: Social saliency prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 4777–4785
23. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: Proc. BMVC. (2011)
24. Chen, C.Y., Grauman, K.: Subjects and their objects: Localizing interactees for a person-centric view of importance. International Journal of Computer Vision (2016) 1–22
25. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
26. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755
27. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2) (2010) 303–338
28. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 248–255
29. : Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. https://github.com/pytorch/pytorch Accessed: 2017-11-03.
30. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: Computer Vision, 2009 IEEE 12th international conference on, IEEE (2009) 2106–2113
31. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: It's written all over your face: Full-face appearance-based gaze estimation. In: Proc. IEEE International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). (2017)
32. Hutman, T., Chela, M.K., Gillespie-Lynch, K., Sigman, M.: Selective visual attention at twelve months: Signs of autism in early social interactions. Journal of autism and developmental disorders **42**(4) (2012) 487–498
33. Baltrušaitis, T., Robinson, P., Morency, L.P.: Openface: an open source facial behavior analysis toolkit. In: Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on, IEEE (2016) 1–10