From Coarse Attention to Fine-Grained Gaze: A Two-stage 3D Fully Convolutional Network for Predicting Eye Gaze in First Person Video

Zehua Zhang[1]
zehzhang@indiana.edu
Sven Bambach[1]
sbambach@indiana.edu
Chen Yu[2]
chenyu@indiana.edu
David J. Crandall[1]
djcran@indiana.edu

[1] School of Informatics, Computing, and Engineering
Indiana University
Bloomington, IN, USA

[2] Psychological and Brain Sciences
Indiana University
Bloomington, IN, USA

Abstract

While predicting where people will look when viewing static scenes has been well-studied, a more challenging problem is to predict gaze within the first-person, ego-centric field of view as people go about daily life. This problem is difficult because where a person looks depends not just on their visual surroundings, but also on the task they have in mind, their own internal state, their past gaze patterns and actions, and non-visual cues (e.g., sounds) that might attract their attention. Using data from head-mounted cameras and eye trackers that record people's egocentric fields of view and gaze, we propose and learn a two-stage 3D fully convolutional network to predict gaze in each egocentric frame. The model estimates a coarse attention region in the first stage, combining it with spatial and temporal features to predict a more precise gaze point in the second stage. We evaluate on a public dataset in which adults carry out specific tasks as well as on a new challenging dataset in which parents and toddlers freely interact with toys and each other, and demonstrate that our model outperforms state-of-the-art baselines.

1 Introduction

We go about our daily lives surrounded by a rich, dynamic, complicated visual world, but can only focus on a small subset of these surroundings at any moment in time [26]. Humans use two mechanisms to direct their visual attention: controlling their field-of-view by moving their head with respect to the world, and controlling the point of foveated vision within the field of view by moving their eye gaze. A critical task for any embodied intelligent system — human or machine — is thus to decide, in real time, where to look. This process is complicated, of course, because the agent must trade off between competing concerns at any moment in time, and these concerns may depend on many factors. Even an everyday action

like crossing the street involves splitting attention between watching for turning vehicles, monitoring oncoming pedestrians, checking that the road ahead is clear, making sure the walk signal has not changed, etc., and this behavior would be quite different if running across a busy highway versus crossing a lonely country road.

Much work in computer vision has considered gaze prediction [8, 24, 29, 31, 32] using combinations of low level features like colors, edges, and texture [31], and higher-level features like scene context [29] and hands [24]. Recent papers use state-of-the-art deep learning models to predict gaze and attention [18, 27, 34] which presumably learn to combine all these cues. But nearly all of this work considers gaze prediction in static photographs or in recorded videos, typically using fixations from multiple human subjects as ground truth data. While this problem is interesting and has many practical applications, it is fundamentally different from the task of deciding where to look *while embodied in a physical environment:* given visual stimulus of a charging lion, for example, one's attention differs dramatically depending on if they are in a movie theater or in the Serengeti. The recent availability of low-cost, lightweight, head-mounted cameras makes it possible to record an approximation of a person's visual field as they go about everyday activities, and portable eye gaze sensors let us record where they are looking within that field of view.

In this paper, we consider the problem of predicting eye gaze in egocentric video captured from head-mounted cameras: given a sequence of several frames, we jointly try to estimate the gaze location point within each, using data from a portable eye gaze tracker as ground truth. A few other papers have studied egocentric gaze prediction [24, 34], but they do so in more controlled contexts (e.g., while subjects are performing a specific task). In contrast, we consider gaze prediction in more free-form scenarios where task information may be unavailable. In particular, we propose a novel deep learning framework that first predicts a coarse attention region, and then combines this attention with spatial and temporal features to predict a precise gaze point. An advantage to this approach is that the coarse region detection could integrate information from non-visual sources, if available — e.g., the task or goal that the camera wearer has in mind — before refining with lower-level visual information. We evaluate on both an existing dataset of controlled activities, showing that our approach is more accurate than the state-of-the-art, and on a new dataset featuring children and adults interacting with objects in an unconstrained toy play scenario. Potential applications of this work include replicating human behavior in robots, safe driving alert systems, attention-driven user interface and advertising, and human-machine interaction.

In summary, our main contributions are: (1) considering a new, challenging scenario for eye gaze prediction; (2) proposing a novel two-stage 3D fully convolutional network that significantly outperforms the state-of-the-art; (3) using a much more challenging dataset for egocentric eye gaze prediction in which subjects engage in free-form activities; and (4) analyzing key components of the model, including the importance of modeling temporal information and the inherent center bias.

2 Related Work

Eye gaze prediction is highly related to work in visual saliency detection, because salient regions tend to attract attention and thus gaze [7]. Most early work [15, 16, 21] takes a bottom-up approach, combining evidence from low-level features [31]. The seminal work of Itti *et al.* [21], for example, combines maps of color, intensity, and edge orientation. Later work gave better performance through better high-level reasoning, implemented using

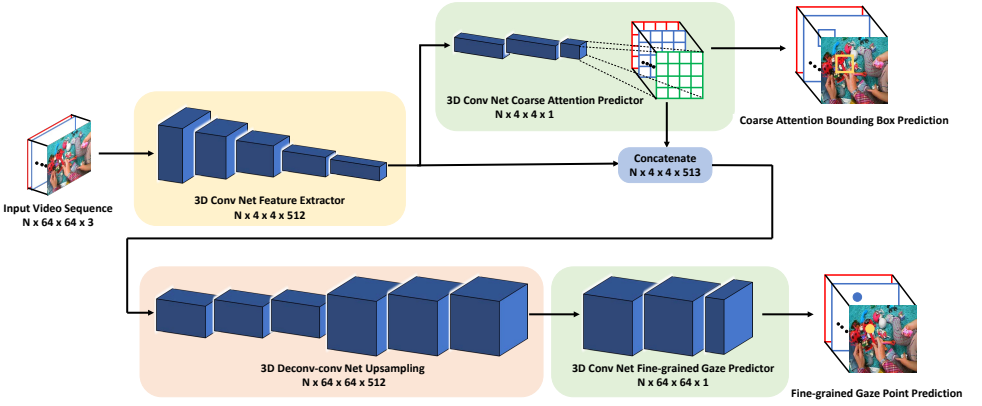


Figure 1: *The architecture of our proposed two-stage 3D fully convolutional network model for eye gaze prediction.* The number below each component indicates its output dimension, where N is the length of the input video sequence.

graph-based [15] and spectral clustering-based models [16], for example. More recently, convolutional neural networks have been used [18, 27], as well as top-down methods such as contextual scene information integration [29] and bottom-up and top-down stream fusing [6]. All of this work is designed for static images, however.

Recently, eye gaze prediction in egocentric videos has been proposed and investigated [3, 24, 32, 34]. Much of this work uses manually-designed features, such as the correlation between head and gaze [3, 32] and the motion and positions of hands [24]; these models may not generalize to other situations when these cues are not available. Zhang *et al*. [34] predict gaze in *future* (unseen) frames by first using a Generative Adversarial Network [13] to "hallucinate" future frames, and then predicting gaze based on spatial and temporal features extracted from these frames. Although their goal is to predict future gaze, as a special case their technique can be used for predicting eye gaze in observable frames.

3 Two-stage 3D Fully Convolutional Gaze Prediction Model

Our goal is to predict where a person is looking within his or her first-person field of view: given an egocentric video of several frames, we wish to label each frame with a single (x, y) coordinate indicating our estimate of their gaze point. There is *exactly one correct answer* per frame in this problem, because the visual stimulus captured in any given frame was only ever experienced and gazed at once, by a single person, in a single fleeting moment. This is significantly different than predicting gaze in recorded scenes, in which a static image can be viewed over a period of time by one or more viewers to collect multiple fixation points.

We propose a two-stage approach to address this problem, by first estimating a coarse region where the eye gaze is likely to lie, and then combining this prediction with other visual and temporal evidence to estimate a more precise gaze point. We use a deep neural network model whose architecture is shown in Figure 1. The input is a sequence of N contiguous frames, which are first fed into a modified C3D [30] network for feature extraction.

A coarse attention predictor uses the extracted features to predict a coarse estimate of the gaze position. That prediction and the extracted feature maps are concatenated and input to an upsampling component, the output of which is used by a fine-grained gaze predictor to produce a more precise gaze location.

3.1 Feature Extraction

Predicting gaze location is difficult, and making reasonable predictions requires combining multiple sources of weak information that depend on the specific context, scenario, and environment. For example, if we know that a subject is preparing breakfast (as in the GTEA [11] and GTEAplus [24] datasets, for example), it is reasonable to assume most gaze will be on food ingredients, so object detection may provide sufficient evidence for accurate gaze prediction. However, since we aim to predict eye gaze in uncontrolled, real-world scenarios, we learn the features automatically to extract as much visual information as possible.

Three-dimensional convolutional networks such as C3D [30] have become popular for capturing spatio-temporal features [22, 23]. To extract features for gaze prediction, we modify C3D by removing the fully-connected layers and the last pooling layer. We also modify the kernel size and the stride of the remaining 4 pooling layers to both be $(1,2,2)$ instead of $(2,2,2)$; this causes the input sequence to be downsampled by a factor of 16 in height and width dimensions, but does not affect the temporal dimension. We leave the architecture of the convolutional layers unchanged so that we can use C3D network weights pretrained on large-scale datasets (e.g., Sports-1M [23]), to simplify our learning. At training and test time, we take a sequence of N RGB images $I_{t,t+N-1}$, resize each frame to 64×64 pixels, and input them to the network, which produces feature $F_{t,t+N-1}$ of dimensionality $N \times 4 \times 4 \times 512$.

3.2 Coarse Attention Prediction

In coarse attention prediction, we divide each video frame into a coarse (e.g., 4×4) grid of cells, and estimate a gaze likelihood distribution over that grid. To do this, we adapt the Fully Convolutional Networks (FCNs) of Long *et al.* [25], who performed dense spatial image labeling (specifically, semantic segmentation) by replacing fully-connected layers with convolutional layers. Here our goal is to predict an approximate attention map for each frame in the input sequence. Since our desired output is exactly a sequence of dense spatial predictions, we design our coarse predictor with 3D convolutional layers inflated from 2D convolutional layers that are transformed from fully-connected layers.

In particular, the input of coarse attention predictor is the set of features extracted from the N frames, $F_{t,t+N-1}$, and the output $C_{t,t+N-1}$ is a sequence of prediction maps with dimensionality $N \times 4 \times 4 \times 1$. We divide each frame into a corresponding 4×4 grid cell. If the eye gaze is within a certain cell, we want the response at the corresponding position in the output map to be 1, and otherwise we want the response to be 0. Since there is exactly one gaze point per image, this can be treated as a classification problem, and flattening the m_{th} output map C_{t+m-1} produces a one-hot vector $Flat(C_{t+m-1})$. Note that predicting which cell contains the true eye gaze is the same as predicting which cell's center is the closest to the gaze. Thus we can use the center of the cell having the highest output from coarse attention predictor as an estimate of the eye gaze point prediction, although the precision of this estimate is obviously limited by the coarseness of the grid.

3.3 Prediction Map Upsampling

To reduce quantization error caused by the coarse attention estimation, we upsample the response maps to have the same spatial resolution as the input image frames. We again follow Long *et al*. [25] and use deconvolution layers to let the network learn the upsampling operation itself [25]. We recombine these maps with the extracted visual features, so that the final gaze estimation can consider evidence from the original frame as well as from the coarse attention prediction's response maps. Since the input of upsampling is the extracted feature along with the coarse prediction, the upsampling is similar to a decoding process; inspired by the decoder of [4], we added two 3D convolutional layers after each deconvolution.

In particular, we concatenate the extracted feature $F_{t,t+N-1}$ and the coarse attention prediction maps $C_{t,t+N-1}$ to obtain a new feature with dimension $N \times 4 \times 4 \times 513$. The upsampling network has two deconv-conv blocks, each of which upsamples its input by a factor of 4 and processes the decoded feature. The output of upsampling, denoted $U_{t,t+N-1}$, has dimension $N \times 64 \times 64 \times 512$, which has the same height and width as the input video frames.

3.4 Fine-grained Gaze Prediction

Like coarse attention prediction, our fine-grained gaze point estimation is a dense labeling task: we aim to generate a likelihood map for each frame and choose the point with the highest probability of gaze. In our model, the fine-grained gaze predictor shares the same network components (number and type of layers, kernel size, stride, number of filters, etc.) as coarse prediction, but the two networks have different weights and are trained separately.

Given the output $U_{t,t+N-1}$ of the upsampling component, the fine-grained gaze predictor produces an output $P_{t,t+N-1}$, which is a sequence of probability maps of dimension $N \times 64 \times 64 \times 1$. We want the m-th output map to be a Gaussian distribution with its center at the gaze point. Note that this is slightly different from coarse attention prediction where the flattened form of the output map is a one-hot vector; this encourages the network to learn that closely neighboring pixels usually have strong correlation (while neighboring coarse cells do not), and improves robustness against slight measurement errors of the gaze trackers.

3.5 Implementation and Training Details

We used a batch normalization layer [19] after each convolutional layer, and ReLU activations for all layers except the outputs of the coarse attention predictor and fine-grained gaze predictor, which use Softmax instead. We tried different input sequence lengths and found $N = 16$ worked best (as we discuss below). Our implementation is based on Keras [10] and Tensorflow [1]. Detailed architecture and other information is available on the project homepage, http://vision.soic.indiana.edu/t3f/. We trained the two stages of our model separately (although it is also possible to train them jointly).

Training the coarse attention predicting stage. We first trained the feature extractor and coarse attention predictor. We initialized the feature extractor with C3D weights [30] pretrained on Sports-1M [23], while the weights of the coarse attention predictor were initialized randomly. We used stochastic gradient descent with learning rate 0.0005, momentum 0.9, decay 0.01 and $L2$ regularizer 10^{-6}. Given a coarse attention output sequence $C_{t,t+N-1}$ and

ground truth $\hat{C}_{t,t+N-1}$, we use a cross entropy loss summed over the N frames,

$$L_{fs} = \sum_{i=t}^{t+N-1} L_{ce}(\hat{C}_i, C_i), \qquad \text{with } L_{ce}(p,q) = -\sum_x \sum_y p(x,y) \log q(x,y), \tag{1}$$

where p and q are 2D maps of true distribution and predicted distribution, and x and y sum over the spatial coordinates.

Training the fine-grained gaze predicting stage. We then train the entire model end-to-end. We initialized the feature extractor and coarse attention predictor with weights obtained in our first-stage training, and randomly initialize the upsampling component and fine-grained gaze predictor. We use stochastic gradient descent with learning rate 0.0001, momentum 0.9, decay 0.01, and $L2$ regularizer 10^{-6}. Given an output map sequence $P_{t,t+N-1}$ and ground truth $\hat{P}_{t,t+N-1}$, our loss function is Kullback-Leibler divergence summed over frames,

$$L_{ss} = \sum_{i=t}^{t+N-1} D_{KL}(\hat{P}_i || P_i), \qquad \text{with } D_{KL}(p,q) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{q(x,y)}, \tag{2}$$

where p and q are 2D maps of true distribution and predicted distribution, and x and y sum over the spatial coordinates of the map.

4 Experiments

We evaluated our technique on two datasets of first-person video collected with gaze tracking, and compared its performance with other state-of-the-art methods and baselines.

4.1 Datasets and Evaluation Metrics

We used two datasets in our evaluation, one collected from adults who were given a specific task, and one collected from parents and toddlers who were freely playing in a toyroom. The **Object Search Task (OST)** dataset of Zhang *et al.* [34] contains 57 video clips in which 55 subjects perform object search and retrieval tasks. Each video lasts about 15 minutes, is recorded at 10 fps with 480×640 resolution, and the horizontal view angle is 60 degrees. We used the same frames and same test and training data split as in their paper to permit a fair direct comparison. In most of these frames, the subjects are performing one specific task (signing a document), so this dataset mimics applications in which the task is known.

In contrast, our new **Adults, Toddlers and Toys (ATT)** dataset provides a more challenging scenario with uncontrolled tasks, interaction between the subject and other people, appearance of many objects including other people's hands, etc. It is based on our dataset collected for studies in developmental psychology [2, 5] and consists of 20 pairs of videos recorded by head-mounted cameras on children and adults as they play together with toys. Each video clip lasts 5 to 10 minutes and has a frame rate of 30 fps. The view angle along the width dimension is 70 degrees. Most videos were taken at a resolution of 480×640, while a small subset have resolution of 480×720 (due to different hardware configurations). In our experiments, we used frames from 18 videos of parents as the training set and 2 videos of parents as the test set. To remove noise, frames in which the eye gaze was outside of the image

Figure 2: *Sample results of our model.* Frames in the first row are from the OST dataset. Frames from the second row to the last row are from the ATT dataset. The big red cross indicates the ground truth gaze point, the little green cross indicates the gaze point prediction, and the blue bounding box indicates the coarse attention cell prediction.

plane or stationary for an unrealistically long time period were discarded, resulting in an average of 10,838 frames remaining in each video. The last four rows of Figure 2 shows sample frames from this dataset. More information is available at http://vision.soic.indiana.edu/t3f/.

Area Under the Curve (AUC) [8] and Average Angular Error (AAE) [28] are two widely used metrics to evaluate the results of eye gaze prediction. However, in our first stage prediction, the expected output is not a saliency map. Though we can take the center of the grid cell with the highest response as an approximate gaze prediction to compute AAE, AUC is not applicable here. For consistency, we use AAE throughout our experiments.

4.2 Results on Object Search Task (OST) Dataset

We first evaluated on the OST dataset, which was recorded from subjects performing specific tasks. Quantitative results in terms of Average Angular Error are reported in Table 1

Method	AAE
Our full model (T3F)	**8.56**
Our coarse predictor	14.32
Our full model with cropping	10.71
Our coarse predictor with cropping	15.03
Deep Future Gaze [34]	10.60
SALICON [18]	13.30
Information Maximization [3]	17.00
Graph-based Visual Saliency (GBVS) [14]	18.80
Itti & Koch [20]	19.00
Adaptive Whitening Salience [12]	22.80
Image Signature [17]	24.20
Saliency using Natural Statistics [33]	25.00

Table 1: *Average Angular Error of our method compared to others,* on the OST dataset. Lower errors are better.

	Number of frames (N)				
	1	2	4	8	16
Coarse attention (CA)	22.86	21.65	20.23	19.08	18.11
Full model (T3F)	19.29	17.85	16.30	15.98	**15.24**

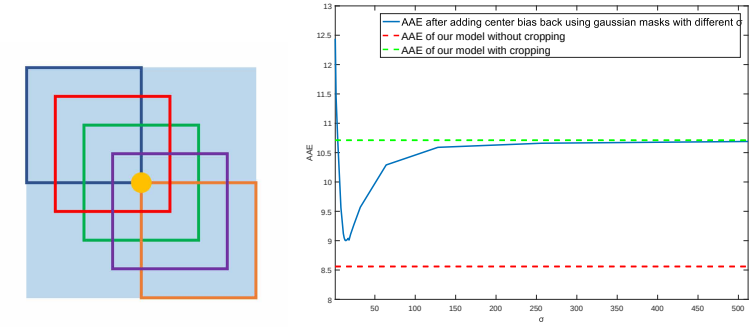
Table 2: *Effect of input sequence length on Average Angular Error of our models,* on the ATT dataset. Lower errors are better.

for several variants of our approach. The table also shows quantitative results for several baseline techniques that predict gaze or saliency. Of these, the only one that was specifically designed for egocentric gaze prediction (Deep Future Gaze [34]) performs best (with an error of 10.60); the other baselines range up to 25.00. Our coarse attention predictor alone scores significantly better than most of these, at 14.32 degrees, while our full model performs better than all by a significant margin, at 8.56. The large error drop between our coarse attention model and the full model shows that our fine-grained gaze prediction stage successfully eliminates the error inherent in coarse-resolution maps. Gaze prediction runs at about 15 fps on our single NVidia Titan X Pascal GPU.

Qualitative results are shown in the first row of Figure 2. Coarse attention tends to find the most task-relevant object or the hands most of the time. Note that this was learned implicitly by the network itself: we do not use an explicit object or hand detector. We also see that fine-grained gaze prediction effectively focuses from a coarse cell to a predicted gaze point. Sometimes the adjustment even makes the final gaze prediction jump out of the coarse attention cell and move towards the ground truth.

4.3 Center Bias in Egocentric Videos

We tried to improve our results by using dataset augmentation, which is of course a very standard technique in vision to discourage overfitting. In particular, we tried cropping image frames by varying amounts to generate additional examples. Surprisingly, however, we found that this caused the performance of our model to actually *decrease*, as Table 1 shows. Our hypothesis is that cropping with scale of at least 2.0 (so that the height and the width



(a) (b)

Figure 3: (a) Illustration of center bias elimination by cropping the images. The yellow point is the eye gaze ground truth and small boxes with different colors indicate randomly cropped images from this image; (b) AAE curve of adding center bias back manually. The blue curve indicates the AAE after adding center bias back using Gaussian masks at the center with different σ, the red dash line indicates the AAE of our model without cropping, and the green dash line indicates the AAE of our model with cropping at the scale of 2.0.

of the cropped image are half of the original image) has the effect of removing the center bias — people tend to gaze at the center of their field of view. This effect is illustrated in Figure 3(a). In other words, if we crop the image randomly at the scale of 2.0, the gaze point can end up anywhere in the cropped image, even though it was in the center of the original.

In order to further study the effect of center bias, we experimented with manually adding center bias back to the predictions of our full model trained with cropping. To do this, we multiplied the output maps with a 2D Gaussian mask having mean at the center. We varied the standard deviation σ of the Gaussian, and plot the AAE values in Figure 3(b). We found that when σ is very small, the error is even greater than that of our model with cropping; a very small σ means adding a very strong center bias. This result shows that our model learns effective visual signals for predicting gaze, instead of just always predicting the center. As σ increases, the AAE begins to drop but is bounded below by the AAE of the full model without cropping, because the center bias is only an approximation of the true bias in egocentric videos. This demonstrates the necessity of allowing the model to learn the bias itself instead of adding it manually. After a certain point, the AAE begins to increase since the strength of the center bias drops as σ increases. Finally, as σ becomes very large, the AAE tends to be equal to that of the model with cropping since it has almost no effect.

4.4 Results on Adults, Toddlers and Toys (ATT) Dataset

Our ATT dataset is very different from OST in that it was recorded in an unstructured environment in which interacting children and parents were freely playing with toys as well as with each other. Qualitative results in the bottom rows of Figure 2 show that our model is robust enough to predict eye gaze in this more challenging dataset. As with OST, the second stage still refines the coarse attention, but the coarse attention (blue rectangles) focuses on more diverse regions because of ATT's complexity: more hands, objects, and people are in view, and we do not know what task the subject has in mind (e.g., which toy the subject likes, how he/she wants to play, whether he/she wants to interact with other people, etc.).

Quantitative results are in Table 2. Our full model has an average error of 15.24 degrees, which is significantly greater than the 8.56 of OST (even considering the differences in field of view between the two cameras), reflecting the difficulty of this dataset.

Table 2 also presents results for different input frame sequence lengths (N), which control how much temporal information the network is able to use. Consistent with intuition, we found that AAE decreases as N increases, but the improvement is capped by the size of the receptive fields. The AAE of the coarse model with $N = 16$ is better than that of the full model with $N = 1$, even though the coarse model introduces significant quantization error. This is because $N = 1$ carries no temporal information, which demonstrates the importance of temporal signals in solving this task. AAE is still decreasing from $N = 8$ to $N = 16$, suggesting that larger sequence lengths may produce even less error. Unfortunately, we were not able to try larger N due to limited GPU memory (on our single NVidia Titan X).

4.5 Ablation Study

To study the benefit of our two-stage structure, we removed the coarse attention predictor and trained the model to directly predict the fine-grained gaze point. On both datasets, we found that the direct model had larger error than our two-stage model (9.02 vs. 8.56 degrees on OST, and 16.09 vs. 15.24 on ATT), which demonstrates the effectiveness of the two-stage structure to improve performance. Interestingly, the direct model also outperforms all other existing baselines (as shown in Table 1), including the state-of-the-art model of Zhang *et al*. [34]. One hypothesis is that unlike their technique (which is designed for predicting gaze in future, unseen frames), ours is able to use information from both previous and future frames to jointly find the best gaze point across all of them. Also, we observe that our direct model with $N = 16$ performs worse than our full model with $N = 8$, despite having more temporal information, again confirming the effectiveness of our two-stage structure.

5 Conclusion

We proposed a novel two-stage 3D fully convolutional network for egocentric eye gaze prediction. We evaluated on two challenging datasets, one of which features uncontrolled tasks, complex objects, and interaction with other people. We demonstrated the capacity of our model to use both spatial and temporal information, including the center bias. Evaluations show our model significantly outperforms other methods. In future study, we plan to evaluate other models on the ATT dataset to compare with our model, and to test our approach on even less constrained datasets. We also plan to incorporate non-visual evidence, such as information about a person's intention, into gaze prediction. Finally, we plan to develop better metrics for evaluation that reflect how gaze detection would work in real-world applications.

Acknowledgments. This work was supported by the National Science Foundation (CAREER IIS-1253549) and the National Institutes of Health (R01 HD074601, R21 EY017843), and the IU Office of the Vice Provost for Research, the College of Arts and Sciences, and the School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project "Learning: Brains, Machines, and Children." We would like to thank Drew Abney, Esther Chen, Steven Elmlinger, Seth Foster, Laura Sloane, Catalina Suarez, Charlene Tay, Yayun Zhang for helping with the collection of the first-person toy play dataset, and Shujon Naha and Satoshi Tsutsui for helpful discussions.

References

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *USENIX Conference on Operating Systems Design and Implementation*, pages 265–283, 2016.

[2] D. H. Abney, H. Karmazyn, L. Smith, and C. Yu. Hand-eye coordination and visual attention in infancy. In *Annual Conference of the Cognitive Science Society (CogSci)*, 2018.

[3] S. O. Ba and J. M. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 33(1):101–116, Jan 2011.

[4] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 39(12):2481–2495, 2017.

[5] Sven Bambach, David Crandall, Linda Smith, and Chen Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Annual Conference of the Cognitive Science Society (CogSci)*, 2016.

[6] A. Borji, D. N. Sihite, and L. Itti. Probabilistic learning of task-specific visual attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[7] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(1):185–207, January 2013.

[8] Ali Borji, Hamed Rezazadegan Tavakoli, Dicky N. Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[9] Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[10] François Chollet, JJ Allaire, et al. R interface to keras. https://github.com/rstudio/keras, 2017.

[11] Alireza Fathi, Yin Li, and James M. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision (ECCV)*, 2012.

[12] Anton Garcia-Diaz, Xose R. Fernandez-Vidal, Xose Manuel Pardo, and Raquel Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, 2012.

[13] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[14] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[15] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems (NIPS)*, pages 545–552, 2007.

[16] X. Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 34(1):194–201, Jan 2012.

[17] Xiaodi Hou, Jonathan Harel, and Christof Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2012.

[18] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning (ICML)*, 2015.

[20] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000.

[21] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[22] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 35(1):221–231, January 2013.

[23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[24] Yin Li, Alireza Fathi, and James M. Rehg. Learning to predict gaze in egocentric video. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[25] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] Michael C Mozer and Mark Sitton. Computational modeling of spatial attention. *Attention*, 9:341–393, 1998.

[27] Junting Pan, Elisa Sayrol, Xavier Giró i Nieto, Kevin McGuinness, and Noel E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–606, 2016.

[28] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[29] Antonio Torralba, Monica S. Castelhano, Aude Oliva, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, 113:2006, 2006.

[30] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. C3D: generic features for video analysis. *arXiv:1412.0767*, 2014.

[31] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97 – 136, 1980.

[32] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Attention prediction in egocentric video using motion and visual saliency. In Yo-Sung Ho, editor, *Advances in Image and Video Technology*, pages 277–288, 2012.

[33] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 2008.

[34] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.