Localizing Novel Attended Objects in Egocentric Views

Shujon Naha[1]
snaha@iu.edu
Md Alimoor Reza[1]
mdreza@iu.edu
Chen Yu[2]
chen.yu@austin.utexas.edu
David Crandall[1]
djcran@indiana.edu

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

[2] Department of Psychology
University of Texas at Austin, USA

Abstract

People have foveated vision and thus are generally able to attend to just a single object within their field of view at a time. Our goal is to learn a model that can automatically identify which object is being attended, given a person's field of view captured by a first person camera. This problem is different from traditional salient object detection because our goal is not to identify all of the salient objects in the scene, but to identify the single object to which the camera wearer is attending. We present a model that learns based on very weak supervision, with just annotations of the label of the class that is attended in each frame, without bounding boxes or other spatial location information. We show that by learning disentangled representations for localization and classification, our model can effectively localize novel attended objects that were never seen during training. We propose a multi-stage *knowledge distillation* strategy to train our generalized localizer model. To the best of our knowledge, our work is the first to explore the problem of learning generalized attended object localization models in egocentric views under weak supervision.

1 Introduction

Camera-enabled wearable devices may soon transform the way people interact with technology. The cameras on these devices can capture people's visual fields on a moment-by-moment basis as they go about their lives. Video from this first-person, egocentric point of view provides a unique perspective of the visual world that is inherently human-centric, giving a level of detail and ubiquity that often exceeds what is possible from environmental, third-person cameras.

It is well known that people have foveated visual systems with high resolution only at one small gaze point (of about 2 degrees [8]) at the center of their field of view. This means that people generally can only pay attention to one object or scene element at a time. Identifying which object is attended or will be attended is of key importance in many first-person computer vision applications, because attention reveals information about the user's

© 2020. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

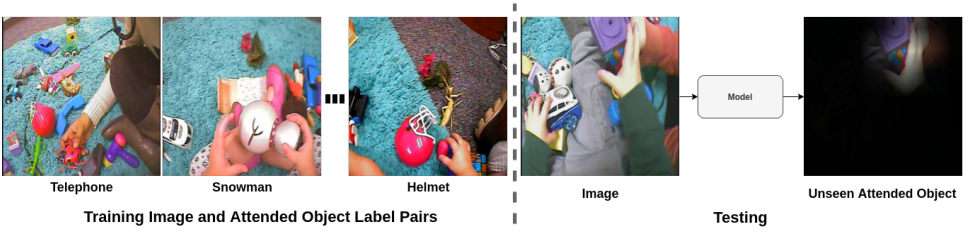

Figure 1: In these images of people's field of views recorded from head-mounted cameras, which objects are they gazing at? We learn a model that can localize the attended object in a cluttered egocentric view, even if the attended object was never seen before in the training data. For example, in the test image, the attended object is the "Purple Box", which was never attended in the training dataset, but our model can still localize it as the attended object. To do this, during training we only use class-label supervision of the attended object — *not* a bounding box or other location information.

activities and intentions on a moment-by-moment basis, without requiring conscious effort. Because of the unique properties of egocentric video, however, estimating attention within the field of view (and without a gaze tracker) is a challenging problem: egocentric views are highly dynamic, with cluttered scenes full of objects from unusual perspectives (Figure 1).

To solve this problem, a system needs to jointly solve two fundamental, inter-related tasks: (1) the "where" problem — which part of the egocentric view the wearer is currently paying attention to, and (2) the "what" problem — which object is being attended. A straightforward solution would be to solve each of these tasks using supervised learning, first training a model to estimate the gaze point based on many labeled training frames, and then training a model to detect and localize all objects in the field of view. More sophisticated approaches [28] could learn joint models to solve both problems simultaneously. But these supervised approaches have two fundamental limitations: (1) they require labeled training data (e.g., with bounding boxes around all objects), which is extremely labor intensive to collect, and (2) they assume that no novel objects — i.e., ones not seen at training time — will ever be attended.

In this paper, we seek to learn a model that, given an egocentric video frame, localizes the object that the camera wearer is visually attending *even if the object was not seen at training time*. We also require only very weak supervision: instead of gaze points and/or object bounding boxes, we simply require, for each training frame, the label (identity) of the object that is visually attended. Thus our learning procedure has to find regularities in the training data to estimate "what" and "where" given only information about "what," and with no spatial information about where the object is located or what it looks like. But we show that an advantage to this approach is that the resulting models are more general: they are able to estimate the location and spatial extent of novel attended objects. Of course, for novel objects, the model will obviously not be able to correctly estimate "what," but we nevertheless expect it to be able to estimate "where."

In more detail, we propose a modular network with two completely separate parts, i.e. a localizer for solving "where," and a classifier for solving "what." Our localizer network first locates the attended region, and then the classifier network identifies the object present in that particular location. By doing so, we essentially disentangle the "where" and "what" problems so that our localizer can find the attention region regardless of the object class present in that region. But training the disentangled model is non-trivial as both the localizer and classifier

Figure 2: Training a localizer network to first locate the attended region and then learn a separate classifier to recognize the object in that region using only the class label can cause ambiguities for both networks. Here in these two input images, the attended objects are the "bed" and the "blue car" respectively but the localizer is looking at the wrong objects i.e., "doll" in both cases. Thus the classifier will get confusing information about the appearance of all three objects and will fail to learn properly which in turn, hurts the localization predictions of the localizer. In contrast, we use knowledge distillation to pretrain the localizer using the same class label supervision, which can help to reduce the initial localization errors.

depend on each other, which makes optimizing these networks a chicken-and-egg problem (Figure 2). Thus to train the disentangled generalized localizer, we follow the widely used concept of knowledge distillation [11] where the learning of a student model is guided by another teacher model. Our teacher model is a regular weakly supervised object localization (WSL) model which uses a single feature representation for solving both the "where" and "what" problems. This model can localize training objects very well but performs poorly for unseen attended objects. We then use this teacher model to initialize our generalized localizer network for class-agnostic attended object localization. Finally, we train our full network (consisting of both the classifier and the pre-trained localizer) end-to-end using only the classification loss. We show that a localizer that exploits temporal information performs significantly better than a single CNN-based localizer network using only the current frame to locate the attended region.

We believe that the generalized attended object localization problem is interesting from both theoretical and practical perspectives. From a cognitive perspective, studying learning systems of this type is interesting because it mimics, in a general sense, the type of learning under weak supervision that human toddlers must do as they learn the words for new objects in cluttered scenes. From a practical technology perspective, this ability to learn from weak supervision and to generalize to new environments will be of significant importance in enabling augmented reality and other camera-enabled devices to operate in challenging real-world environments.

2 Related Work

We briefly discuss relevant previous works on active object recognition, eye-gaze (attention) localization, and weakly-supervised saliency prediction. Our work is also reminiscent of early computer vision work in foveated vision systems [2, 4, 6].

Active Object. Pirsiavash et al. [20] distinguished between active and non-active objects from egocentric videos. Zhou and colleagues proposed a cascade neural network to predict hand segmentation maps and active objects [29]. A spatio-temporal binning approach is proposed in the work of McCandless and Grauman [19] to localize manipulated objects. Zhang et al. [28] predicted the bounding box of the currently attended object by using a

modified SSD [18] network and utilizing both spatial and temporal information. But all of these papers require strong supervision while learning to localize active objects, whereas we use only the attended object's class label (without bounding boxes or other location information).

Eye Gaze. Yamada et al. [24] combined bottom-up saliency with head motion to predict eye gaze location. Generative Adversarial Networks and 3D-CNNs were explored by Zhang et al. [26] to predict gaze location on future frames. Zhang et al. [27] used a two-stage 3D CNN to generate coarse attention maps and then refine them to predict eye gaze location. However, none of these works explored the relationship between attended object class and eye gaze location under a very weak supervision – i.e., instead of requiring gaze points and/or object bounding boxes, only exploiting the label of the attended object.

Weakly Supervised Saliency Prediction. Wang et al. [23] proposed a two-stage training strategy for training a saliency prediction model with only class label supervision. Hsu et al. [13] proposed a similar strategy to learn to generate foreground-background maps using only salient object class label supervision. Zeng et al. [25] used both class labels and captions to produce saliency maps. Unlike these approaches which considered all salient objects as foreground, we must localize the *single object* that the camera wearer is attending among all objects in cluttered first-person views.

3 Our Approach

Our approach trains the generalized Localizer network L in three stages: (1) training a teacher weakly-supervised object localization (WSL) model, (2) propagating the localization knowledge from the teacher model to the generalized localizer L, and (3) further training the localizer L using a separate classifier network C for better, more general, attended object localization. Figure 3 depicts the three training stages along with our complete model.

3.1 Stage I: Training the Specialized WSL Network

In the first stage, we train a regular weakly supervised object localization (WSL) model. We use WildCat [7] in our implementation, but any WSL model could be used as long as it only uses class label supervision for training, and produces class activation maps for the predicted object class. The class activation map of the predicted object class can be used to find the location of the attended object. To correctly predict the attended object class, the WSL model needs to activate the region of the attended object while suppressing activations for the other objects in view. This implies the WSL model needs to solve both the "where" and "what" problems for accurate prediction. Since the localizer and classifier are entangled in the WSL model, its localizer gets more direct supervision during training compared to our disentangled localizer L (Figure 3). As a result, the WSL model can learn the locations and appearances of the specific attended objects *present in the training set* with less ambiguity compared to our disentangled model. We train the WSL model using a cross-entropy loss as follows:

$$loss_{stage1} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\in p}\mathbb{I}[Y^i = j]\log P^i_j, \tag{1}$$

where Y^i is the ground truth attended object class label for instance i, p is the set of the training classes, and P^i is the softmax probability scores generated after global average pooling of the

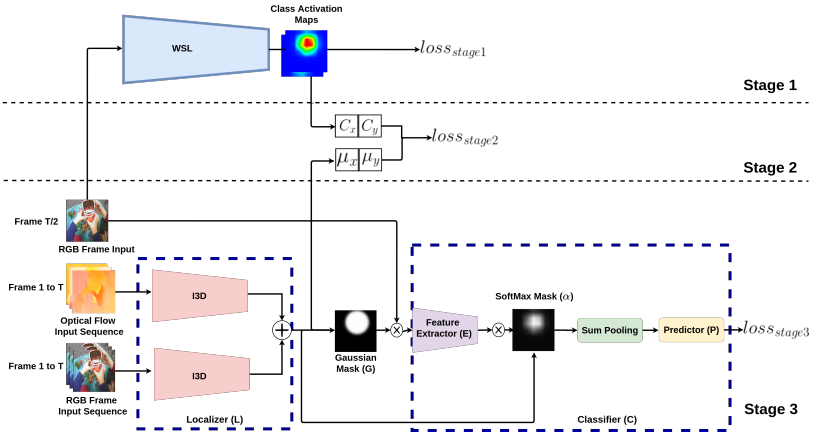

Figure 3: We train our generalized attended object localization model (L) in three stages. Stage I trains an existing Weakly Supervised Localization (WSL) model using only the class label supervision. In Stage II, the WSL model works as a guide to initialize the localizer network L of our model. The <C_x, C_y> and <μ_x, μ_y> are the predicted attended object location centers from the WSL model and our generalized localizer model L, respectively. In Stage III, both the classifier (consisting of feature extractor (E) and predictor (P)) and the localizer networks are trained using only class label supervision.

class activation maps produced by WSL. $\mathbb{I}[.]$ is an indicator function that produces 1 for true and 0 for false argument. Although WildCat is not specifically designed for locating attended objects from the egocentric view, we found that it could learn to often successfully localize the attended object region, presumably cueing on regularities about objects that tend to be attended – the largest in view, most center in view, closest to the hand, etc.

3.2 Stage II: Knowledge Distillation To The Generalized Localizer

As the classifier and the localizer are tightly bound together in the WSL model, they fail to localize attended objects that were not seen in the training data. We thus only use the WSL model as a teacher to help guide our generalized localizer network L. We consider the center of mass of the class activation map generated by the WSL model for the predicted object class as the position of the attended object. Let us denote the 2D location of the center as <C_x, C_y>.

In the second stage of our training, we consider <C_x, C_y> as pseudo ground truth and we use them to pre-train our generalized localizer network L. The localizer network consists of two separate I3D networks [5] for optical flow and RGB inputs, each pretrained on Kinetics [13]. It takes a sequence of RGB frames and its corresponding optical flows, and then computes an attention map for the attended object in the middle (current) frame. This way, the network sees both the past and future frames while predicting the location of the attended object in the current frame.

Let $s_1 \in \mathbb{R}^{T\times H\times W\times 3}$ and $s_2 \in \mathbb{R}^{T\times H\times W\times 2}$ be T frames of RGB and optical flow, respectively. After we pass s_1 and s_2 through the localizer network, we extract two attention maps $l_{rgb} \in \mathbb{R}^{h\times w}$ and $l_{flow} \in \mathbb{R}^{h\times w}$. Each attention map is generated by first passing the last convolutional feature of the corresponding I3D network through a 3D convolution layer to reduce the number of channels to one. Temporal pooling is then applied on the final feature map to

convert it to a 2D attention map. We fuse these two feature maps using a summation, $l = l_{rgb} + l_{flow}$. This fused feature map l is used as the final location map of the our localizer network, which we then upscale to $H \times W$ using bilinear interpolation. We apply a 2D Softmax on l to generate attention map α and then follow a similar approach as in [14] to find the 2D center location $<\mu_x, \mu_y>$ of α which we consider as the attended object location,

$$\begin{aligned} \alpha &= Softmax(l), \\ a_x = \Sigma_{y=1}^{H} \alpha, a_y &= \Sigma_{x=1}^{W} \alpha, \\ \mu_x = \frac{\Sigma_{x=1}^{W} x * \exp(a_x)}{\Sigma_{x=1}^{W} \exp(a_x)}, \mu_y &= \frac{\Sigma_{y=1}^{H} y * \exp(a_y)}{\Sigma_{y=1}^{H} \exp(a_y)}. \end{aligned} \tag{2}$$

We then train the localizer network using an L_1 loss,

$$loss_{stage2} = |C_x - \mu_x| + |C_y - \mu_y|. \tag{3}$$

It may seem that training the localizer L using $loss_{stage2}$ should be sufficient for generalized localization of attended objects. However, experimentally we found that training L using only $loss_{stage2}$ overfits the model to the objects in the training images. To avoid this overfitting, we train L with $loss_{stage2}$ only for very few epochs, so that this stage serves primarily as an initialization or pre-training step for the localization network L.

3.3 Stage III: Training The Localizer Using A Disentangled Classifier

In the third stage, we further train the localizer network L by attaching a separate classifier network on top of it (Figure 3 bottom right) to solve both the "what" and "where" problems together. This new classifier network consists of a feature extractor E and a predictor P. Our hypothesis is that the localizer now must solve a more general and semantically meaningful problem, instead of just directly optimizing a loss for location prediction (such as $loss_{stage2}$). In this stage, the localizer network L is still responsible for finding the location of the attended object, but receives supervision from the classifier network in terms of verification of the attended object class in the predicted region. The classifier network recognizes the object in the location predicted by the localizer network and uses the ground truth object label for calculating the training loss. Since the L network is already initialized from stage II, it generally predicts correct attended locations, allowing the classifier to learn a good appearance model for the attended object classes. The classifier can then further improve the accuracy of the pretrained localizer from stage II by propagating the error gradients from the classifier network to the localizer network L in stage III.

In order to accomplish this goal, we first generate a 2D Gaussian map $G = (\mu_x, \mu_y, \sigma)$ centered at the softmax attention map α. The classifier receives the middle frame of the input sequence masked by the predicted Gaussian map G from the localizer network. This essentially allows the classifier network to focus on only the attended region.

Let I be the RGB input of the middle frame. We generate a masked version of I through an elementwise multiplication with G, $I' = G \odot I$. The masked image I' is then passed through the feature extractor part E of the classifier network to generate a feature representation F. Then we convert F to a vector embedding v using weighted sum-pooling by considering α as the weight,

$$v = \sum_j \alpha_j * F_j, \tag{4}$$

where j is the pixel location. The embedding v is then passed through the predictor P of the classifier network to identify the object in the attended location. We use a loss that jointly optimizes the localizer and the classifier networks,

$$loss_{stage3} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\in p}\mathbb{I}[Y^i = j]\log P(v)^i_j. \tag{5}$$

4 Experiments

4.1 Dataset

We train our model using the *Toy Room* [1, 3] dataset. This dataset is particularly suitable for evaluating our approach as it is an egocentric dataset with a ground truth class label and bounding box location of the currently attended object in each frame. The dataset was collected from toddlers playing with their parents in a room with 24 toys. The children wore a head-mounted forward-facing camera, and a gaze tracker to measure attention. For training, we used the dataset's annotations indicating the label of the attended object in each frame; again, we do not use localization data (e.g., gaze point location or attended object bounding box) during training. For testing, the model was given only the input image frames with no additional annotations, and we used the attended object bounding boxes to evaluate our localization accuracy. For both training and evaluating our model, each instance is a sequence of 16 RGB and corresponding optical flow frames where the 8th RGB frame is considered as the current frame. Our training data includes only 19 object classes from the 24 toys, while the remaining 5 object classes are used in the test data. Thus our model has to localize attended objects which were not seen during training time. In total, we used 16,390 training instances and 7,328 test instances in our experiments. During inference, we discard the classifier and only use the localizer model for locating the unseen attended objects.

To test the generalizability of our model, we also tried training our model on the Toy Room dataset but testing on a completely different dataset, GTEA Gaze+ [9, 16]. GTEA Gaze+ contains 24 *fps* videos of adults performing various cooking-related tasks while wearing a head-mounted camera. This dataset contains gaze location ground truth for each frame, which we use for our evaluation. Similar to the Toy Room dataset, each instance in this dataset contains a sequence of 16 RGB and Optical Flow frames. We use this dataset only for evaluation.

4.2 Implementation Details

For each of the training stages, we use Stochastic Gradient Descent (SGD) for training with a momentum value of 0.9 and weight decay 10^{-7}. We use a mini-batch size of 32 in each iteration. For stage I, we use the Wildcat [7] model as our WSL network. We train the model for 20 epochs using $loss_{stage1}$ with learning rate 0.0001. For stage II, we only train L for two epochs with learning rate 0.01 to avoid overfitting on the training data. For stage III, we use a ResNet101 [10] model pretrained on Imagenet as the feature extractor E, and a sequence of two fully-connected layers as the predictor P for the classification model. We use learning rates 0.01 and 0.001 for P and E, respectively, and trained the model for 30 epochs. We used PWC-Net [22] for generating the optical flow inputs. Finally, for generating the Gaussian map G, we use a fixed standard deviation of 0.4.

Method	Train	Test
PicaNet [17]	54.32	64.51
DSS [12]	58.30	66.06
WILDCAT [7] (Stage I)	96.92	68.70
ResNet Localizer (Stage II)	81.92	70.03
Ours (Stage II)	82.75	77.66
ResNet Localizer (Stage III)	94.52	74.33
Ours (Stage III)	91.03	**82.53**

Table 1: Evaluation of generalized attended object localization accuracy on unseen test attended objects in the ToyRoom dataset. We show the localization accuracy for both the train and test data.

4.3 Baselines

We consider several baseline models to evaluate the effectiveness of our localization model. As our generalized attended object localization problem has some similarities with the class-agnostic salient object localization problem, we consider two deep saliency models, PiCANet [17] and DSS [12], which are trained using ground truth saliency maps. We also use Wildcat [7] trained on the Toy Room dataset as a baseline to show its performance on unseen attended objects. For the saliency models and Wildcat, we consider the center of mass of the predicted saliency map and the class activation map as the attended object location. For a final baseline, we replace our localizer with a single ResNet101 [10] model (ResNet Localizer) to show the effectiveness of using temporal data for attended object localization.

4.4 Point Localization

In this experiment, we evaluate the performance of our localizer L (Figure 3) for locating attending objects irrespective of their class. We use a point localization measure similar to that of [7]. In this setting, an estimated location (μ_x, μ_y) is considered accurate if it is within 15 pixels of the ground truth bounding box of the attended object. As shown in Table 1, the results suggest that our problem is significantly harder than predicting saliency as both the saliency models perform much worse than the other approaches. This is probably because the egocentric view often contains multiple salient objects, but only one of them was actually attended. The saliency models seem to highlight all the salient regions instead of only the attended object region, which is also evident from Figure 4. We observe that Wildcat performs extremely well for localizing the attended objects seen in the training dataset, but, unsurprisingly, it performs much worse for the unseen objects in the test data. During inference time we use the class activation map of the predicted class from Wildcat for localization, which gives ambiguous information for the unseen objects. The ResNet localizer baseline performs significantly better than Wildcat, which shows that disentangling the localizer and classifier improves the generalization capability of the localizer model.

Finally, our full model significantly outperforms all the baselines on the test dataset. The accuracy on the training data is much higher for Wildcat, which implies that it learns class-specific biases in the dataset. However, the high training object localization accuracy of Wildcat also justifies our usage of $<C_x, C_y>$ as pseudo ground truth for the attended object location at stage II. Training a few epochs with $loss_{stage2}$ and further training the model in stage III subsequently helps to train our localizer network without those biases and improves

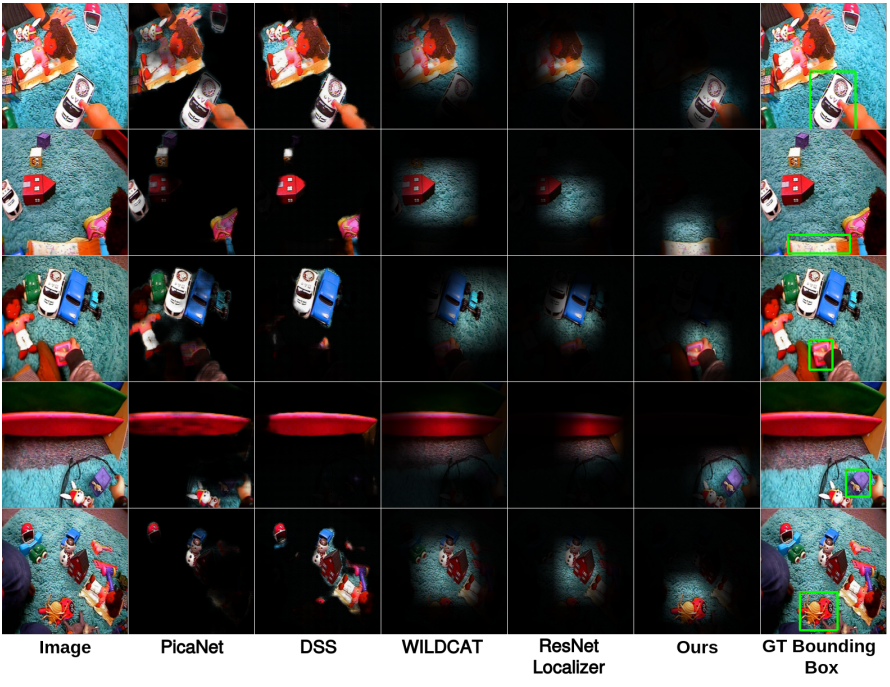

Figure 4: Qualitative results on the Toy Room dataset. For the ResNet classifier and our model, we visualize the predicted gaussian mask G.

Method	PiCANet [17]	DSS [12]	ResNet Localizer (Stage III)	Ours (Stage III)
AAE	13.31	12.73	10.83	**10.62**

Table 2: Evaluation of eye gaze prediction error on GTEA Gaze+ (**lower is better**). The ResNet Localizer and our model are trained using weak supervision on the ToyRoom dataset. The other two models are trained with ground truth saliency maps from different datasets.

the localizer's generalization capability. Also, we see that the accuracy of the final stage III model is significantly higher than that of stage II, which demonstrates that our training of the localizer using the classification loss in stage III is important. Our model also performs significantly better than the ResNet localizer baseline, indicating that temporal information improves the attended object localization accuracy. We also tried to train our model without using Wildcat as initialization (i.e., skipping stages I and II), and found the results varied widely (from 52.18% to 81.23%) on the test dataset across different runs. This indicates that a good initialization is required to properly train the disentangled localizer model.

4.5 Gaze Prediction

To measure the ability of our model to generalize across different settings, we took the localizer model trained on the Toy Room dataset and applied it on the GTEA Gaze+ dataset, and evaluated on eye gaze location prediction. This task is extremely challenging as the settings of these two datasets are very different: Toy Room was collected from toddlers

playing with toys with their parents, while GTEA Gaze+ was collected by adult subjects wearing head-mounted cameras while performing cooking activities in the kitchen. We use the widely-used Average Angular Error (AAE) [21] metric for measuring performance, which indicates the distance between the estimated attention point (μ_x, μ_y) and the ground truth gaze point. As baselines, we consider the saliency models and the ResNet localizer for predicting the eye gaze location. The results are shown in Table 2. The AAE of our model is 10.62 which is significantly lower than both the saliency models and the ResNet localizer network. This suggests that our model learned a more general concept of attended regions compared to the baseline models.

5 Conclusion

In this paper, we introduce the novel problem of generalized localization of attended objects from egocentric videos only using class label supervision during training. We show that a specialized weakly supervised model, which uses the same feature representation for both classification and localization of the attended object, can locate attended objects seen at training time but fails to localize novel attended objects. We propose a generalized localizer model and a multi-stage knowledge distillation strategy, and show that our approach can effectively disentangle and propagate localization knowledge from the specialized model to the generalized localizer model. We also show that our model can be applied to different egocentric settings other than the training dataset, which indicates its effectiveness as a generalized attended region localization model.

Acknowledgements. This work was supported in part by the National Institute of Child Health and Human Development (R01HD074601 and R01HD093792), the National Science Foundation (CAREER IIS-1253549), and the Indiana University Office of the Vice Provost for Research, the College of Arts and Sciences, and the Luddy School of Informatics, Computing, and Engineering through the Emerging Areas of Research Project *Learning: Brains, Machines and Children*.

References

[1] 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.

[2] Dana H Ballard. Animate vision. *Artificial intelligence*, 48(1):57–86, 1991.

[3] S. Bambach, D. Crandall, L. Smith, and C. Yu. Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Annual Conference of the Cognitive Science Society (CogSci)*, 2016.

[4] Marc Bolduc and Martin D Levine. A review of biologically motivated space-variant data reduction models for robotic vision. *Computer vision and image understanding*, 69(2):170–184, 1998.

[5] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[6] Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4):497–523, 1990.

[7] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 642–651, 2017.

[8] Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013.

[9] A. Fathi, Y. Li, and JM. Rehg. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision (ECCV)*, pages 314–327, 2012.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

[13] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised salient object detection by learning a classifier-driven map generator. *IEEE Transactions on Image Processing*, 28(11): 5435–5449, 2019.

[14] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 4020–4031, 2018.

[15] 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.

[16] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 287–295, 2015.

[17] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018.

[18] 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.

[19] T. McCandless and K. Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. *British Machine Vision Conference (BMVC)*, 2:3, 2013.

[20] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. *International Conference on Computer Vision and Pattern Recognition*, pages 2847–2854, 2012.

[21] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1153–1160, 2013.

[22] D. Sun, X. Yang, M. Liu, and J. Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. *International Conference on Computer Vision and Pattern Recognition*, 2018.

[23] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 136–145, 2017.

[24] K. Yamada, Y. Sugano, T. Okabe, Y. Sato, A. Sugimoto, and K. Hiraki. Attention prediction in egocentric video using motion and visual saliency. *Pacific-Rim Symposium on Image and Video Technology*, pages 277–288, 2011.

[25] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6074–6083, 2019.

[26] M. Zhang, K. Ma, J. Lim, Q. Zhao, and J. Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4372–4381, 2017.

[27] Z. Zhang, S. Bambach, D. Crandall, and C. Yu. From coarse attention to fine-grained gaze: A two-stage 3d fully convolutional network for predicting eye gaze in first person video. In *British Machine Vision Conference (BMVC)*, 2018.

[28] Zehua Zhang, Chen Yu, and David Crandall. A self validation network for object-level human attention estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[29] Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian. Cascaded interactional targeting network for egocentric video analysis. *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1904–1913, 2016.